The role of structural pleiotropy in the retention of protein complexes after gene duplication

Mémoire

Angel Fernando Cisneros Caballero

Maîtrise en biochimie - avec mémoire
Maître ès sciences (M. Sc.)

Québec, Canada

© Angel Fernando Cisneros Caballero, 2019

The role of structural pleiotropy in the retention of protein complexes after gene duplication

Master's thesis

Angel Fernando Cisneros Caballero

Under the supervision of:
Christian R. Landry

Résumé

La duplication de gènes est l'un des plus importants mécanismes évolutifs pour la génération de diversité fonctionelle. Lorsqu'un gène est dupliqué, la nouvelle copie partage toutes ses fonctions avec la copie ancestrale car elles encodent pour des protéines identiques. Donc, les deux protéines, appelées paralogues, auront le même réseau d'interactions physiques protéine-protéine. Cependant, dans le cas de la duplication des gènes qui codent des protéines qui interagissent avec elles-mêmes (homomères), la nouvelle protéine interagira aussi avec la copie ancestrale, ce qui introduit une nouvelle interaction (heteromère) (Kaltenegger and Ober, 2015; Pereira-Leal et al., 2007). Puisque ces interactions peuvent avoir des différents motifs de rétention et de fonction (Ashenberg et al., 2011; Baker et al., 2013; Boncoeur et al., 2012; Bridgham et al., 2008), il est important de mieux comprendre comment ces états sont atteints et quelles forces évolutives les favorisent. Dans ce memoire, je cible ces questions avec des simulations *in silico* de l'évolution des protéines suite à la duplication de gènes en travaillant avec des structures crystallographiques de haute qualité, provenant de la *Protein Data Bank* (Berman et al., 2000; Dey et al., 2018). Les simulations montrent que les sous-unités et interfaces partagées entraînent une forte corrélation entre les trajectoires évolutives de ces complexes. Ainsi, les simulations prédisent que la préservation de seulement les deux homomères ou seulement l'hétéromère ne devrait pas être fréquente. Toutefois, la simulation qui applique la sélection seulement sur un homomère montre que l'homomère neutre est destabilisé plus rapidement que l'hétéromère neutre. Nous avons comparé ces prédictions avec des résultats expérimentaux du réseau d'interactions protéine-protéine de la levure. Comme suggéré par les simulations, les patrons d'interactions les plus fréquents ont été la formation des trois complexes (deux homomères et un hétéromère) ou la formation de seulement un homomère. Les patrons correspondants à deux homomères sans hétéromères ou un hétéromère sans homomères sont rares. Nos résultats démontrent l'extension de l'hétéromérisation entre paralogues dans le réseau d'interactions physiques protéine-protéine de la levure, les mécanismes sous-jacents et ses implications.

Abstract

Gene duplication is one of the most important evolutionary mechanisms for the generation of functional diversity. When a gene is duplicated, the new copy shares all of the ancestral copy's functions because they encode identical proteins. Therefore, the two proteins, called paralogs, will have the same protein-protein interaction network. However, in the case of the duplication of genes encoding proteins that self-interact (homomers), the new protein will also interact with the ancestral copy, introducing a novel interaction (heteromer) (Kaltenegger and Ober, 2015; Pereira-Leal et al., 2007). As these interactions can have different retention and functional patterns (Ashenberg et al., 2011; Baker et al., 2013; Boncoeur et al., 2012; Bridgham et al., 2008), it is important to understand better how these states are reached and what evolutionary forces favor each of them. In this thesis, I approach these questions by means of *in silico* simulations of protein evolution after gene duplication by working with high-quality crystal structures from the Protein Data Bank (Berman et al., 2000; Dey et al., 2018). The simulations show that the shared subunits and interfaces lead to these complexes having highly correlated evolutionary trajectories. Thus, the simulations predict that the preservation of only the two homomers or only the heteromer is not likely to happen often. Nevertheless, simulating evolution with selection on only one homomer shows that the neutral homomer is destabilized faster than the neutral heteromer. We compared these predictions against experimental results from the yeast protein-protein interaction network. As suggested by the simulations, the most abundant interaction patterns were either the formation of all three complexes (two homomers and one heteromer) or the formation of only one homomer, with motifs corresponding to two homomers without a heteromer or a heteromer without homomers being rare. Our results highlight the extent of heteromerization between paralogs in the yeast protein-protein interaction network, the underlying mechanisms, and its implications.

Contents

Résumé . . . . . **ii**

Abstract . . . . . **iii**

List of Figures . . . . . **v**

List of Abbreviations and Acronyms . . . . . **vi**

Acknowledgements . . . . . **viii**

Foreword . . . . . **ix**

Introduction **1**

1 Chapter 1 . . . . . **8**
- Résumé 8
- Abstract 9
- Introduction 10
- Results 12
- Discussion 22
- Material and methods 25
- Author contributions 36
- Acknowledgements 37
- Competing interests 37

2 Discussion **38**

Conclusion **42**

Bibliography **43**

Annex **56**
- Supplementary text 56
  - Comparison of PCA results with previous studies 56

List of Supplementary Figures **57**

List of Figures

1 Mutations in paralogous proteins originating from an ancestral homomer are likely to have pleiotropic effects on each other's function due to their physical association 10
2 Homomers and heteromers of paralogs are frequent in the yeast protein interaction network 13
3 The loss of heteromerization between paralogs leads to greater functional divergence . . 16
4 Selection to maintain homomers also maintains heteromers. 18
5 Mutations have pleiotropic effects on homomers and heteromers. 19
6 The loss of heteromerization between paralogs may result from regulatory divergence. . 21

List of Abbreviations and Acronyms

- INDEL = insertion/deletion
- NMR = nuclear magnetic resonance
- cryo-EM = cryo-electron microscopy
- PDB = Protein Data Bank
- WGD = whole-genome duplication
- SSD = small-scale duplication
- GO = gene ontology
- HM = homomer
- HET = heteromer
- HM&HET = homomer and heteromer
- PCA = Protein-fragment Complementation Assay
- 2D = duplicated both by SSD and WGD
- DHFR = dihydrofolate reductase
- NAT = nourseothricin
- HygB = hygromycin B
- EDTA = ethylenediaminetetraacetic acid
- MTX = methotrexate
- DMSO = dimethylsulfoxide
- FRET = fluorescence resonance energy transfer
- BiFC = bimolecular fluorescence complementation
- PCR = polymerase chain reaction
- dNTP = deoxynucleotide triphosphate
- YGOB = Yeast Gene Order Browser

To my mom, my dad, and my younger brother, who made me the person I am today.

"If I have seen further than others, it is by standing upon the shoulders of giants"

Isaac Newton, 1675

Acknowledgements

This thesis completes a great adventure for me that started two years ago. It has been a very exciting journey that has taken me to discover a lot of new ideas, people, cultures, and places. Thus, this would be a good moment to thank everyone that has helped me through all this time.

It all starts with Carmen, Miguel, and Mauricio: my mom, my dad, and my younger brother. Their unconditional support has been a great source of motivation for me to keep moving forward. Through all these years, I have learned a lot of them, from my first words in Spanish and in English to how to live. Being away from them has been a challenge at times, but knowing they are there for me is a great source of inspiration.

I want to thank the rest of my family and my closest friends: Juan Pablo Terrazas, Rodrigo Ledesma, Gilberto Nava, Erick Cervantes, Patricia Hernández, Ricardo Morales, Fernando Ramírez, Daniel Rodríguez, Miguel Angel González, Salvador Paniagua, Rubén Pedrero, Jorge Peña, Rodrigo Delgado, Sergio Villicaña, Santiago Ceballos, David Lerner, Manuel Cortés, and Elisa Fernández. Though at different times, they have all been critical to me for their support on both a personal and a professional level.

New journeys bring new friends. I want to thank Carla Bautista, Andrés Carrillo, Natalia Sinuco, Rafael Ireta, Verônica Alves, Camilo Castañeda, Ulyses Caram, Daniela Pérez, Honey Jain, Johanna Benndorf, and Karolina Jackowska for making me feel at home far away from my country, even when on the other side of the Atlantic.

As for my scientific career, it is important for me to thank my teachers. Their hard work has also been an inspiration to me, and I am glad to say that I have learned from them. Special thanks to Jonathan Muñoz, Patricia Segura, Benito Antón, Susana Martiñón, Edgar Mixcoha, Ricardo Hernández, Claudia Rangel, and Christian Landry. They all believed in me at different stages of my career and gave me platforms to work harder and reach further. In particular, working with Christian for the past two years has been an amazing experience in terms of the leadership, organization, communication, and scientific skills that he shows everyday.

Another great part of this work has been the contributions of my coworkers at the Landrylab and collaborators. Our laboratory has provided a great environment to discuss scientific ideas that gives us reasons to be proud about and push harder with our collective successes. Another great collaborator I want to thank is Joseph Marsh, who welcomed me in his laboratory for an internship. Similarly, I would like to thank my evaluators: Rong Shi and Patrick Lagüe, whose feedback helped me improve the quality of this thesis.

Last but not least, all of this is possible thanks to the funding organisms that support my research. I want to thank Mitacs; the Mexican Association for Cooperation and International Development (AMEXCID) and the Ministry of Education and Higher Education of Québec (MEES); Hydro-Québec and the Université Laval Foundation; the Québec Network for Research on Protein Function, Engineering, and Applications (PROTEO); and the Company of Biologists for their funding contributions.

Foreword

This Master's thesis contains the results of the main project I carried out at Christian Landry's laboratory from September 2017 to April 2019. Such results were included in a journal article titled "The role of structural pleiotropy and regulatory evolution in the retention of heteromers of paralogs", which was initially submitted to *eLife* on March 26th, 2019. Revisions were received on April 26th, 2019, after which a revised version of the manuscript was submitted. Ultimately, it was accepted by *eLife* and published on August 27th, 2019. Axelle Marchant and I were co-first authors for this article, and it also includes work done by Alexandre K. Dubé, Isabelle Gagnon-Arsenault, Diana Ascencio, Honey A. Jain, Simon Aubé, Chris Eberlein, Daniel Evans-Yamamoto, Nozomu Yachie, and Christian R. Landry. I am thankful to all the coauthors for accepting the use of this article in my Master's thesis.

The version of the article I included here has a few differences with respect to the one that was published by *eLife*:

- I removed the references to the supplementary tables and the key resources table because they are too wide to be included properly in the annex of this thesis. The supplementary tables contain the dataset compiled from all the sources used, oligonucleotide sequences used, additional details on the RNA sequencing analysis, the list of PDB structures used for each analysis, and counts of substitutions accepted in each of the simulations. They are available online in the journal's website.
- In the "Pleiotropy contributes to the maintenance of heteromers" section, I added a short comment on how Dey et al. (2018) determined high-quality structures in their analysis.
- In the "Identification of interfaces" section, I clarified that the definitions of contacting and nearby residues were taken from Keskin et al. (2005); Tsai et al. (1996). I also added a short comment on why nearby residues were defined using a cutoff of 6 Å.
- The supplementary text is included as part of the annex.
- Supplementary figures are labeled with an "S" and their number rather than the *eLife* system in which they are supplements of the main figures. They are shown in the same order as in the original article.
- The author contributions and acknowledgements are listed immediately after the materials and methods section.

As stated in the author contributions of the published article, I participated in the design of the computational part of the study, performed the *in silico* simulations of protein complex evolution and the analysis of protein structures, and analyzed the results. I wrote the manuscript together with Axelle Marchant and Christian R. Landry with input from all authors.

During my Master's studies, I also performed other analyses together with Honey Ashok Jain, worked with Luciano Gama Braga from Sabine Elowe's laboratory at the CHUL, and visited Joseph Marsh's laboratory at the University of Edinburgh. This allowed me to explore other topics and ideas related to protein evolution. The collaboration with Luciano Gama Braga and Sabine Elowe resulted in a manuscript available as a preprint at `https://www.biorxiv.org/content/10.1101/733378v1`

that we also submitted for publication. Though I do not discuss these results in this thesis, I comment briefly on the discussions and work done with Honey Ashok Jain and Joseph Marsh as part of the discussion.

Introduction

Protein complexes in biology

Proteins are the main workhorses of the cell. They are biomolecules formed from sequences of 20 different kinds of amino acids. They carry out a great diversity of functions, including the formation of biological structures, metabolic processes, and signaling functions (Boeckmann et al., 2005; Pandey et al., 2017; Scott and Pawson, 2009). Understanding how proteins get access to this wide diversity of functions has been one of the most important questions for modern biology because it has implications for diverse fields, such as medicine and evolutionary biology. Protein functions are conserved across evolution because they have an influence on traits. As such, knowing how they carry out specific tasks helps us understand better how new traits are acquired (Abrusán and Marsh, 2018; DePristo et al., 2005; Ferrada and Wagner, 2008; Kachroo et al., 2015; Payne and Wagner, 2019). As a result, the disruption of protein function could lead to alterations in such traits, and ultimately, diseases (Sahni et al., 2015; Schuster-Böckler and Bateman, 2008).

Research has pointed to two critical properties of proteins that drive their function: protein folding and the formation of protein complexes (Janin et al., 2008; Marsh and Teichmann, 2015; Pandey et al., 2017; Scott and Pawson, 2009; Vidal et al., 2011). Protein folding refers to the three-dimensional configuration of proteins in the cellular space. Current models of protein folding predict it to happen in an orderly fashion by means of local interactions between amino acids (Englander and Mayne, 2014; Maity et al., 2005). A protein's fold then becomes important when studying how it interacts with its environment, most notably with other proteins. Protein complexes are physical associations of proteins, which can either be formed by identical or different chains and are found in a great diversity of stoichiometries (Ahnert et al., 2015). The formation of complexes has been pointed out to be critical for the function of proteins with respect to bringing components of metabolic pathways together and carrying out basic functions in the cell, such as the production of other proteins, because the formation of some active sites depends on complex assembly (Abrusán and Marsh, 2018; Korostelev et al., 2006). Whereas a protein's fold determines how proteins interact with their environment, there are proteins, called chaperones, that contribute to the correct folding of other proteins by interacting with them (Mayer, 2010), so folding and interactions have an impact on each other.

Even though protein complexes are constrained by having to maintain their function, they also change throughout evolution. Consequences of the evolution of protein complexes are the gain of new subunits (Levy et al., 2008; Marsh and Teichmann, 2014), which leads to differences in the distribution of types of protein complexes in the different kingdoms of life (Lynch, 2012) and the appearance of new functions (Boncoeur et al., 2012; Bridgham et al., 2008). In the light of these observations, I am interested in modelling the evolution of protein complexes. However, in order to model it, we must take a closer look at where proteins come from, the processes through which they evolve, and their properties.

Where do proteins come from and how do they evolve?

Proteins are encoded by genes, that is, DNA sequences that organisms pass on to future generations. This process is carried out by two successive steps: transcription of the genetic information from DNA to RNA, and translation, in which proteins are produced based on the RNA sequence (Brenner et al., 1961; Gros et al., 1961; Jacob and Monod, 1961). Thus, protein sequences reflect the underlying DNA sequences of the genes that encode them, so studying genome evolution can yield insight into protein evolution.

Genomes evolve through different processes that lead to changes in the properties that allow proteins to perform their function, such as the efficiency of folding or binding to other molecules (DePristo et al., 2005). These effects ultimately have an effect on the fitness and the chances of survival of organisms, as they could lead to the appearance of a beneficial function or compromise a fundamental function. With time, natural selection will tend to reduce the frequency of genomic variants that are less fit, while the fitter variants will tend to increase their frequency in populations (Crow and Kimura, 1970; Orr, 2009). These processes include recombination, point mutations, and gene duplication. Recombination is a process by which homologous DNA molecules exchange genetic information, leading to a high variability among the individuals in a population (Kaniecki et al., 2018). Point mutations, on the other hand, are changes in a position in the DNA sequence that can have a wide variety of effects. They can be silent if they do not result in a change in the protein sequence due to the underlying architecture of the genetic code (Dufton, 1983; Firnberg and Ostermeier, 2013), but they can still be either beneficial or harmful (Parmley and Hurst, 2007). Other types of point mutations, termed INDELs because they insert or delete nucleotides from the DNA sequence, lead to shifts in the reading frame and result in changes to the protein sequence downstream of the mutation (Lin et al., 2017). Finally, gene duplication results in the birth of a new copy of a gene, which initially leads to an increase in the concentration of a protein but can lead to many different outcomes (Ohno, 1970).

Though these processes drive protein sequence evolution, there are several constraints that shape it. Mutation rates have been described as biased towards a higher genomic AT content, which, given the structure of the genetic code, would support the appearance of, on average, more hydrophobic residues (Bastolla et al., 2004; Dufton, 1983; Hershberg and Petrov, 2010). To some extent, these biases are reflected in the relative abundance of amino acids in the proteome, suggesting that evolution by neutral expectations can lead to proteomes with similar amino acid contents. Nevertheless, there are some deviations from those expectations, which would be imposed by natural selection (Dufton, 1983). This bias could happen due to the specific properties of different regions within the same protein chain. For example, proteins are exposed to a polar solvent with a largely crowded environment of different kinds of molecules with which they can interact. As such, hydrophobic residues play important roles for protein folding and in the binding interfaces of protein complexes, whereas more polar residues are exposed to the solvent (Levy et al., 2012). Still, contacts between protein chains must be restricted to very specific regions because the introduction of hydrophobic residues in other parts of the protein surface could lead to protein aggregation, which can compromise protein function (Garcia-Seisdedos

et al., 2017). Indeed, diseases such as sickle cell anemia and Alzheimer's disease have been linked to protein aggregation (Aguzzi and O'Connor, 2010; Garcia-Seisdedos et al., 2017).

Gene duplication

I am particularly interested in the evolution of proteins after gene duplication because it is one of the main drivers of functional innovation. Sizable percentages of modern genomes arose through gene duplication, ranging from 17% in bacterial species to 65% in *Arabidopsis thaliana* (Zhang, 2003). In fact, whole-genome duplications (WGDs) have been documented several times across evolution. Some broadly described cases are that of *Saccharomyces cerevisiae* (Kellis et al., 2004; Marcet-Houben and Gabaldón, 2015; Wolfe, 2015), the two rounds of WGDs in the common ancestor of vertebrates (Dehal and Boore, 2005), and the several rounds of WGDs of the *Arabidopsis* (Bomblies and Madlung, 2014) and *Paramecium* lineages (Aury et al., 2006).

As gene duplication initially results in an increase of the production of a given protein, it can have effects on fitness. For example, the high copy numbers of ribosomal genes that arose from duplication have been conserved across evolution because their products are required at high levels (Sugino and Innan, 2006). Duplicate genes, or paralogs, can also show distinct retention patterns in different organisms due to the presence or absence of selective pressure on those functions. Such is the case of the reduction of the number of functional olfactory receptor genes in human with respect to mice, which has been attributed to a higher selective pressure on the sense of smell in mice (Rouquier et al., 2000). Other consequences of genome duplication could be the bridging of species barriers, leading to hybridization between species as a way to share genetic material and produce lineages that could have access to greater fitness because of the rapid production of genetic and phenotypic diversity (Arnold and Martin, 2010; Charron et al., 2019).

There are two classes of gene duplication events: whole-genome duplications (WGD) and small-scale duplications (SSD). WGDs occur due to errors in cell division, both in mitosis and meiosis. These events lead to daughter cells that have more copies of the whole set of chromosomes than the progenitor cells (Otto and Whitton, 2000). SSDs occur through unequal crossing over during recombination events and retrotransposition (Reams and Roth, 2015; Xing et al., 2006; Zhang, 2003). Unequal crossing over can happen when repetitive sequences flanking a gene engage in recombination with a slight shift, leading to the exchange of genetic material between two different loci such that one of the chromosomes gets two adjacent copies of the locus (Reams and Roth, 2015; Zhang, 2003). On the other hand, retrotransposons are DNA sequences that are transcribed to RNA molecules that are then used as templates for retrotranscription to produce more DNA copies of themselves that are later integrated elsewhere in the genome. These sequences are capable of carrying over other genetic elements with them, including genes, which has resulted in duplication events (Xing et al., 2006; Zhang, 2003).

Despite the positive effects on fitness of some gene duplications, others can have a negative effects. A mechanism for this negative effect is the dosage balance hypothesis, which proposes that proteins are expressed in levels required for their participation in protein complexes. In this sense, it is necessary to distinguish between WGDs and SSDs because they have different consequences on the balance of protein subunits (Hakes et al., 2007). WGD events imply an increase in the concentration of every protein, so they do not cause a change in the relative abundance of the subunits of protein complexes. SSD events, on the other hand, imply that only the concentration of one of the subunits increases, which results in an excess of protein chains that might not participate in complexes (Edger and Pires, 2009; Papp et al., 2003). The consequences of each type of duplication lead to differences in the types of genes that are retained after each of these and the similarity of the retained paralogs, as evidenced by sequence identity, gene ontology (GO) terms, and shared interactions (Guan et al., 2007; Hakes et al., 2007). As per lost duplicate genes, they often become pseudogenes, often described as "molecular fossils" in the sense that they once were genes but lost their expression or their function (Zhang, 2003).

As some of the consequences of gene duplication can be explained through its effects on protein complexes, it is important to study them together. Following gene duplication, the new duplicate protein will share all of its interactions with the ancestral protein (Diss et al., 2017; Pereira-Leal et al., 2007). As there are now two genetic loci that produce the same protein, there is some degree of redundancy in their function, which results in the two paralogs buffering each other's loss (DeLuna et al., 2008; Hsiao and Vitkup, 2008). A potential mechanism for compensation could be the upregulation of the maintained paralog (DeLuna et al., 2010) so that it can replace the lost one in protein complexes (Gagnon-Arsenault et al., 2013). However, the two paralogs diverge with time as they accumulate mutations. Potential results of this are the uncoupling of the ancestral protein's functions as to make these functions independent from each other. Thus, the two paralogs might split the ancestral functions, a process called subfunctionalization (Baker et al., 2013; Force et al., 1999; Lynch and Force, 2000); or one of them might gain a new function, a process called neofunctionalization (Boncoeur et al., 2012; Bridgham et al., 2008). Some studies discuss the capability of proteins to accumulate neutral variants can prepare them later to acquire new functions, given that they maintain their proper fold, and how these new functions might increase their frequency in populations (Innan and Kondrashov, 2010; Rastogi and Liberles, 2005; Teufel et al., 2019, 2016). Others discuss that regulatory evolution could lead to differences in the expression patterns of paralogs, which has an impact on their retention (Braasch et al., 2016; Gout and Lynch, 2015; Lan and Pritchard, 2016; Lien et al., 2016; Sandve et al., 2018) and their involvement in tissue-selective diseases (Barshir et al., 2018).

A particular case that has been less studied is the duplication of self-interacting, or homomeric, proteins. In this specific case, the two duplicate copies will share the capability of self-interaction. Nevertheless, as they are undistinguishable at first, they will also interact with one another, leading to the formation of heteromers (Diss et al., 2017; Kaltenegger and Ober, 2015; Pereira-Leal et al., 2007). This phenomenon is called paralog interference because it results in a competition of the two proteins for their participation in protein complexes. Pereira-Leal et al. (2007) described the duplication of homomeric proteins as a way to introduce asymmetry, which they pointed out to promote the diversification of interaction partners. Notably, they also observed that heteromers of paralogs

are more often formed by contacts between paralogs than by contacts between identical subunits, suggesting that duplication influences the assembly of complexes. Since then, other studies have yielded insight into the variety of ways in which the function of protein complexes can evolve following paralog interference. Baker et al. (2013) identified that a pair of duplicated MADS-box transcription factors split the regulatory targets of the ancestral protein. In *Kluyveromyces lactis*, a species of yeast in which the transcription factor has not been duplicated, homodimers of the MADS-box transcription factors regulate the expression of genes involved in arginine metabolism and the Mat-α mating type. On the other hand, in *Saccharomyces cerevisiae*, in which the transcription factor has been duplicated, the homodimer of one of the duplicates regulates arginine metabolism while their heterodimer regulates the expression of Mat-α mating type genes. Other families of transcription factors also exhibit this pattern of different targets being regulated by homomers and heteromers (Amoutzias et al., 2008). Boncoeur et al. (2012) described how a pair of duplicates, *patA* and *patB*, evolved to provide a new function. The heteromer formed by these proteins is a bacterial multidrug transporter, but they are only capable of providing this function when they are expressed together and heteromerize. Ashenberg et al. (2011) showed that the duplicate histidine kinases EnvZ and RstB have evolved to form only homomers in *Escherichia coli* by means of competition assays. They suggest that such specificity might be achieved by the accumulation of neutral mutations and the later fixation of critical destabilizing mutations. Finally, they explain that the loss of the heterodimer could be attributed to potential deleterious effects of crosstalk between histidine kinases, as they tend to autophosphorylate upon homomerization and this could extend to phosphorylation upon heteromerization.

These results have prompted other studies on the extent and importance of heteromeric interactions between paralogs in protein-protein interaction networks. Diss et al. (2017) showed that when some proteins are deleted, their paralogs lose some of their interactions, which suggests that these paralogs might depend on each other to participate in protein complexes and carry out their functions. An observation that supports this model of paralog dependency is that the deletion of proteins is sometimes associated to the degradation of their paralogs (DeLuna et al., 2010). Interestingly, paralog dependency can extend directly to protein function, as in the case of the Fam20A and Fam20C pair of duplicates. Fam20A on its own is an inactive kinase, whereas Fam20C is an active kinase on its own. However, the heteromeric association of Fam20C with the Fam20A pseudokinase was shown to have an increased kinase activity with respect to the Fam20C homodimer (Cui et al., 2015). Generalizing this model of paralog dependency leads to the prediction that the deletion of proteins that interact with their paralogs should have more deleterious fitness effects than the deletion of non-interacting paralogs, which was confirmed recently by analyses of large-scale CRISPR-Cas9 deletion datasets (Dandage and Landry, 2019).

Considering the implications of heteromerization for the functional evolution of interfering paralogs, my objective is to model how sequence divergence affects the retention of protein interactions from a biophysical point of view. By means of this model, we would be able to study if there are evolutionary constraints for the loss of heteromers and the possibility of the development of specificity. As such, we must look at the biophysical properties of protein complexes and the available computational tools that will allow us to model evolution *in silico*.

Protein structure and biophysics

The field of protein biophysics has received considerable attention in recent years as a way to address the long-standing questions about protein folding and the assembly of protein complexes. Protein structures are usually solved by means of different experimental techniques, such as X-ray crystallography, nuclear magnetic resonance (NMR), and cryo-electron microscopy (cryo-EM), with each of these being favored for different objectives. While X-ray crystallography is the main workhorse of structural biology because of the atomic resolution of the structures it can generate (Shi, 2014), NMR can provide useful information about protein dynamics (Kleckner and Foster, 2011), and cryo-EM can be used to look at larger protein complexes that are not easily crystallized (Costa et al., 2017), with recent developments pushing cryo-EM towards the range of atomic resolutions of X-ray crystallography (Shi, 2014). These techniques have contributed to the resolution of around 150 000 protein structures, which are deposited on databases such as the Protein Data Bank (PDB) (Berman et al., 2000), making them an invaluable resource for structural biology. However, care must be taken when selecting structures because the techniques used to solve them are not perfect and sometimes are subject to erroneous interpretations. Therefore, different surveys of quality of the structural models have been carried out, both based on the reinterpretation of results (Wlodawer et al., 2018) and the evolutionary conservation of assemblies, which would support the biological significance of the structures (Dey et al., 2018).

The availability of structural data has led to the development of increasingly complex models on protein structure. These include early amino acid contact potentials (Miyazawa and Jernigan, 1996); energy functions that look at the thermodynamic contributions of different kinds of interactions (hydrogen bonding, van der Waals forces, hydrophobic interactions, etc.), such as FoldX (Guerois et al., 2002) and Rosetta (Alford et al., 2017); and molecular dynamics simulations that integrate protein dynamics into the model (Hollingsworth and Dror, 2018). These models provide information about protein chain stability, that is, a measure of how favorable it is for a protein to fold the right way, and of the binding energy of complexes, which explains how favorable it is for proteins to assemble into complexes (Bastolla et al., 2017; DePristo et al., 2005; Liberles et al., 2012), on different levels. As such, the choice of energy function depends on the objective and the computational resources available. For example, molecular dynamics simulations are useful to address questions about protein dynamics and conformational changes with a higher level of detail but often require supercomputers due to the complexity of the calculations that must be carried out. In turn, other energy functions are more suitable for evolutionary models because they provide reasonable estimations of the energetic values in a shorter time frame, allowing a more extensive sampling of the sequence space.

Though based on prior knowledge of proteins, these models have led to the confirmation of hypotheses through simulations and the proposal of new ideas on the properties of proteins. An early model suggested that symmetrical homomeric assemblies implied that any mutation in such complexes that creates an interaction between residues from different chains would create them at more than one spot. Thus, a prediction of this model was that these kinds of complexes would be more easily acquired than heteromeric assemblies and thus more abundant (Monod et al., 1965). This hypothesis

was confirmed later by means of simulations that showed that even random peptides have an inherent propensity to interact with other copies of themselves (Lukatsky et al., 2007, 2006). André et al. (2008) later observed that in simulations that selected for complex formation, symmetric complexes became dominant over time, even if they started out as a very small fraction of the total complexes. As a result, they concluded that the most favorable assemblies of homomers out of random samples were symmetric, which is in line with analyses of quaternary structures (Ahnert et al., 2015; Bergendahl and Marsh, 2017).

Other models have used fitness functions based on energetic terms to describe the evolution of protein complexes. Kachroo et al. (2015) showed that some proteins that have similar functions in different species, termed orthologs, are capable of complementing each other when the native one is deleted and replaced with the one from the other species. Using simulations, they explained that complementarity between complexes is maintained because the mutations that distinguish the orthologous proteins have much lower effects on the binding energy and stability of the protein chains than random mutations. As such, those would be the product of neutral variation that preserves the protein's properties and function. Teufel et al. (2019) used a similar evolutionary model to study the evolution of a heterodimeric complex after the gene that encodes one of its subunits is duplicated. They concluded that, depending on the selection scenario, the evolutionary trajectories of the simulations can lead to quite different outcomes. In particular, when selection acts to preserve binding to one of the duplicates but acts against the other, the model predicts that the two resulting complexes could be quickly destabilized at first with subsequent recovery of the favored interface. This model highlights how a heteromeric interface could become specific to only one of the two duplicates.

As per my interests, these last two models represent powerful tools to model the evolution of duplicate genes. While Teufel et al. (2019) looked at the evolution of paralogs after gene duplication, this model is most suitable for heteromeric interactions that are not subject to paralog interference. Therefore, the model I am interested in working with is a model with paralog interference, in which the ancestral protein forms a homomer and is then duplicated. As discussed above, this gives birth to a second homomer and a heteromer of the duplicates. Considering that selection on these complexes would have to distinguish between a protein chain binding to itself and binding to a highly similar copy of itself, the evolution of these complexes might be subject to different evolutionary constraints to those modelled by Teufel et al. (2019). By working with different scenarios, I will be able to evaluate whether selection can favor or not a particular kind of complex over the other. As such, my working hypothesis will be that different selection scenarios will lead to different outcomes, with selection on one of the complexes maintaining it while destabilizing the others.

1 Chapter 1

Marchant, A.*, Cisneros, A. F.*, Dubé, A., Gagnon-Arsenault, I., Ascencio, D., Jain, H., Aubé, S., Eberlein, C., Evans-Yamamoto, D., Yachie, N., Landry, C. R., (2019). The role of structural pleiotropy and regulatory evolution in the retention of heteromers of paralogs. *eLife*, 8, 46754.

* Denotes equal contributions

Résumé

La duplication des gènes favorise l'évolution de nouvelles fonctions. La duplication des gènes codant des protéines homomériques entraîne la formation d'homomères et d'hétéromères de paralogues, ce qui crée de nouveaux complexes suite à un seul événement de duplication. La perte de ces hétéromères peut être requise pour que les deux paralogues évoluent des fonctions indépendantes. En utilisant la levure comme modèle, nous démontrons que l'hétéromérisation est fréquente entre les homomères dupliqués et est corrélée avec la similarité fonctionnelle entre paralogues. En utilisant, l'évolution *in silico*, nous démontrons que pour les homomères et les hétéromères qui partagent des interfaces de liaison, les mutations chez un paralogue peuvent avoir des effets structuraux pléiotropiques sur les deux interactions, ce qui entraîne des réponses fortement corrélées à la séléction. En conséquence, l'hétéromérsiation pourrait être préservé indirectement à cause de la sélection sur la rétention des homomères, ce qui ralentit la divergence fonctionnelle entre paralogues. Nous suggérons que les paralogues peuvent supérer l'obstacle de la pléiotropie structurelle grâce à l'évolution régulatrice aux niveaux transcriptionnel et post-traductionnel.

Abstract

Gene duplication is a driver of the evolution of new functions. The duplication of genes encoding homomeric proteins leads to the formation of homomers and heteromers of paralogs, creating new complexes after a single duplication event. The loss of these heteromers may be required for the two paralogs to evolve independent functions. Using yeast as a model, we find that heteromerization is frequent among duplicated homomers and correlates with functional similarity between paralogs. Using *in silico* evolution, we show that for homomers and heteromers sharing binding interfaces, mutations in one paralog can have structural pleiotropic effects on both interactions, resulting in highly correlated responses of the complexes to selection. Therefore, heteromerization could be preserved indirectly due to selection for the maintenance of homomers, thus slowing down functional divergence between paralogs. We suggest that paralogs can overcome the obstacle of structural pleiotropy by regulatory evolution at the transcriptional and post-translational levels.

Introduction

Proteins assemble into molecular complexes that perform and regulate structural, metabolic and signaling functions (Janin et al., 2008; Marsh and Teichmann, 2015; Pandey et al., 2017; Scott and Pawson, 2009; Vidal et al., 2011; Wan et al., 2015). The assembly of complexes is necessary for protein function and thus constrains the sequence space available for protein evolution. One direct consequence of protein-protein interactions (PPIs) is that a mutation in a given gene can have pleiotropic effects on other genes' functions through physical associations. Therefore, to understand how genes and cellular systems evolve, we need to consider physical interactions as part of the environmental factors shaping a gene's evolutionary trajectory (Landry et al., 2013; Levy et al., 2012).

A context in which PPIs and pleiotropy may be particularly important is during the evolution of new genes after duplication events (Amoutzias et al., 2008; Baker et al., 2013; Diss et al., 2017; Kaltenegger and Ober, 2015). The molecular environment of a protein in this context includes its paralog if the duplicates derived from an ancestral gene encoding a self-interacting protein (homomer) (Figure 1). In this case, mutations in one paralog could have functional consequences for the other copy because the duplication of a homomeric protein leads not only to the formation of two homomers but also to a new heteromer (Figure 1) (Pereira-Leal et al., 2007; Wagner, 2003). We refer to these complexes as homomers (HMs) and heteromers of paralogs (HETs).

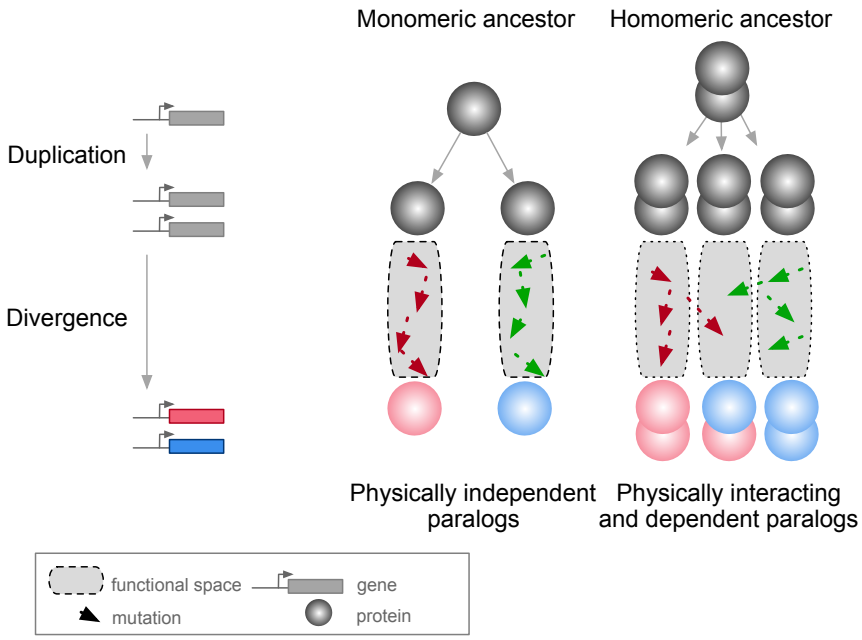

Figure 1: Mutations in paralogous proteins originating from an ancestral homomer are likely to have pleiotropic effects on each other's function due to their physical association.
Gene duplication leads to physically interacting paralogs when they derive from an ancestral homomeric protein. The evolutionary fates of the physically associated paralogs tend to be interdependent because mutations in one gene can impact on the function of the other copy through heteromerization.

Paralogs originating from HMs are physically associated as HETs when they arise. Subsequent evo-

lution can lead to the maintenance or the loss of these HETs. Consequently, paralogs that maintained the ability to form HETs have often evolved new functional relationships (Amoutzias et al., 2008; Baker et al., 2013; Kaltenegger and Ober, 2015). Examples include a paralog degenerating and becoming a repressor of the other copy (Bridgham et al., 2008), pairs of paralogs that split the functions of the ancestral HM between one of the HMs and the HET (Baker et al., 2013), that cross-stabilize and that thus need each other to perform their function (Diss et al., 2017), or that evolved a new function together as a HET (Boncoeur et al., 2012). However, there are also paralogs that do form HMs but that have lost the ability to form HETs through evolution. Among these are duplicated histidine kinases (Ashenberg et al., 2011) and many heat-shock proteins (Hochberg et al., 2018). For the majority of HETs, we do not know what novel functions, if any, contribute to their maintenance.

Therefore, one important question to examine is: what are the evolutionary forces at work for the maintenance or the disruption of HETs arising from HMs? Previous studies suggest that if a paralog pair maintains its ability to form HMs, it is very likely to maintain the HET complex as well (Pereira-Leal et al., 2007). For instance, Lukatsky et al. (2007) showed that proteins tend to intrinsically interact with themselves and that negative selection may be needed to disrupt HMs. Since nascent paralogs are identical just after duplication, they would tend to maintain a high propensity to assemble with each other. Hence, the two paralogs would form both HMs and HETs until the emergence of mutations that specifically destabilize one or the other (Ashenberg et al., 2011; Hochberg et al., 2018). In addition, the rate at which the HET is lost may depend on epistasis since it may cause mutations to be more or less disruptive together for the HET than they are individually for the HMs (Diss and Lehner, 2018; Starr and Thornton, 2016). Here, we hypothesize that the association of paralogs forming HETs acts as a constraint that may slow down the functional divergence of paralogs by making mutations on one paralog affect the function of the other.

Previous studies have shown that HMs are enriched in eukaryotic PPI networks (Lynch, 2012; Pereira-Leal et al., 2007). However, the extent to which paralogs interact with each other has not been comprehensively quantified in any species. We therefore analyze the physical assembly of HETs exhaustively in a eukaryotic interactome by integrating data from the literature and by performing a large-scale PPI screening experiment. Then, using functional data analysis, we examine the consequences of losing HET formation for paralogs forming HMs. We perform *in silico* evolution experiments to study whether the molecular pleiotropy of mutations, caused by shared binding interfaces between HM and HET complexes, could contribute to maintain interactions between paralogs originating from ancestral HMs. We show that selection to maintain HMs alone may be sufficient to prevent the loss of HETs. Finally, we find that regulatory evolution, either at the level of gene transcription or protein localization, may relieve the pleiotropic constraints maintaining the interaction of paralogous proteins.

Results

Homomers among singletons and paralogs in the yeast PPI network

We first examined the extent of homomerization across the yeast proteome (see dataset in Materials and methods and the supplementary text) for two classes of paralogs, those that are small-scale duplicates (SSDs) and those that are whole-genome duplicates (WGDs). We considered these two sets separately because they may have been retained through different mechanisms (see below). The dataset for this analysis, which includes previously reported PPIs and novel DHFR Protein-fragment Complementation Assay experiments (referred to as PCA, see Materials and methods and supplementary text), covers 2521 singletons, 2547 SSDs, 866 WGDs and 136 genes that are both SSDs and WGDs (henceforth referred to as 2D). We find that among the 6070 tested yeast proteins, 1944 (32%) form HMs, which agrees with previous estimates from crystal structures (Lynch, 2012). The proportion of HMs among singletons (n = 630, 25%) is lower than for all duplicates: SSDs (n = 980, 38%, p-value<2.0e-16), WGDs (n = 283, 33%, p-value=1.6e-05) and 2D (n = 51, 38%, p-value=1.7e-03) (Figure 2A).

Although a large number of PPIs have been previously reported in *Saccharomyces cerevisiae*, it is possible that the frequency of HMs is slightly underestimated because they were not systematically and comprehensively tested (see Materials and methods). Another reason could be that some interactions were not detected due to low expression levels. We measured mRNA abundance in cells grown in PCA conditions and used available yeast protein abundance data (Wang et al., 2012) to test this possibility. As previously observed (Celaj et al., 2017; Freschi et al., 2013), we found a correlation between PCA signal and expression level, both at the level of mRNA and protein abundance (Spearman's r = 0.33, p-value = 3.5e-13 and Spearman's r = 0.46, p-value < 2.2e-16 respectively). When focusing only on previously reported HMs, we also observed both correlations (Spearman's r = 0.37, p-value = 3.9e-08 and Spearman's r = 0.38, p-value = 6.0e-08 respectively). The association between PCA signal and expression translates into a roughly two-fold increase in the probability of HM detection when mRNA levels change by one order of magnitude (Figure S1A). We also generally detected stronger PCA signal for the HM of the most expressed paralog of a pair, confirming the effect of expression on our ability to detect PPIs (Figure S1B). Finally, we found that HMs reported in the literature but not detected by PCA have on average lower expression levels (Figure S1B-C). We therefore conclude that some HMs (and also HETs) remain undetected because of low expression levels.

The overrepresentation of HMs among duplicates was initially observed for human paralogs (Pérez-Bercoff et al., 2010). One potential mechanism to explain this finding is that homomeric proteins are more likely to be maintained as pairs after duplication because they might become dependent on each other for their stability that is enhanced through the formation of HET (Diss et al., 2017). Another explanation is that proteins forming HMs could be expressed at higher levels and thus more easily detected, as shown above. High expression levels are also associated with a greater long term probability of genes to persist after duplication (Gout et al., 2010; Gout and Lynch, 2015). We indeed

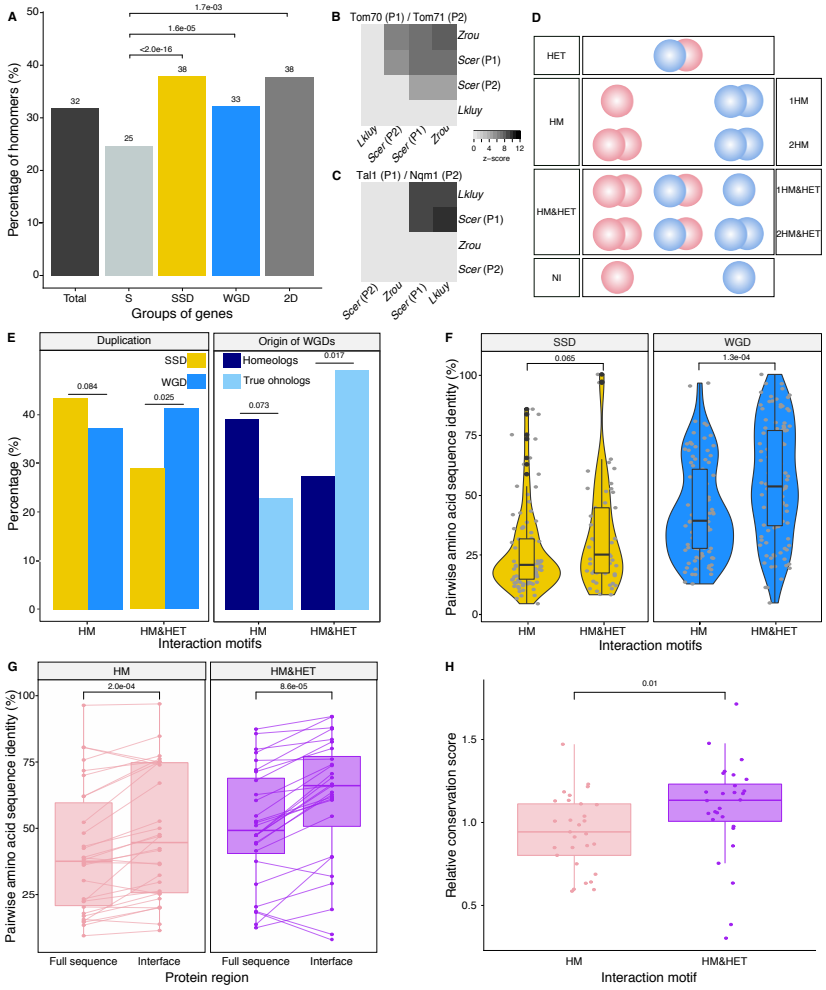

Figure 2: Homomers and heteromers of paralogs are frequent in the yeast protein interaction network. **(A)** The percentage of homomeric proteins in S. cerevisiae varies among singletons (S, n = 2521 tested), small-scale duplicates (SSDs, n = 2547 tested), whole-genome duplicates (WGDs, n = 866 tested) and genes duplicated by the two types of duplication (2D, n = 136 tested) (global Chi-square test: p-value<2.2e-16). Each category is compared with the singletons using a Fisher's exact test. P-values are reported on the graph. **(B and C)** Interactions between *S. cerevisiae* paralogs and pre-whole-genome duplication orthologs using DHFR PCA. The gray tone shows the PCA signal intensity converted to z-scores. Experiments were performed in *S. cerevisiae*. Interactions are tested among: **(B)** *S. cerevisiae* (Scer) paralogs Tom70 (P1) and Tom71 (P2) and their orthologs in *Lachancea kluyveri* (Lkluy, SAKL0E10956g) and in *Zygosaccharomyces rouxii* (Zrou, ZYRO0G06512g) and **(C)** *S. cerevisiae* paralogs Tal1 (P1) and Nqm1 (P2) and their orthologs in *L. kluyveri* (Lkluy, SAKL0B04642g) and in *Z. rouxii* (Zrou, ZYRO0A12914g). **(D)** Paralogs show six interaction motifs that we grouped in four categories according to their patterns. HET pairs show heteromers only. HM pairs show at least one homomer (one for 1HM or two for 2HM). HM&HET pairs show at least one homomer (one for 1HM&HET or two for 2HM&HET) and the heteromer. NI (non-interacting) pairs show no interaction. We focused our analysis on pairs derived from an ancestral HM, which we assume are pairs showing the HM and HM&HET motifs. **(E)** Percentage of HM and HM&HET among SSDs (202 pairs considered, yellow) and WGDs (260 pairs considered, blue) (left panel), homeologs that originated from inter-species hybridization (47 pairs annotated and considered, dark blue) (right panel) and true ohnologs from the whole-genome duplication (82 pairs annotated and considered, light blue). P-values are from Fisher's exact tests. **(F)** Percentage of pairwise amino acid sequence identity between paralogs for HM and HM&HET motifs for SSDs and WGDs. P-values are from Wilcoxon tests. **(G)** Pairwise amino acid sequence identity for the full sequences of paralogs and their binding interfaces for the two motifs HM and HM&HET. P-values are from paired Wilcoxon tests. **(H)** Relative conservation scores for the two motifs of paralogs. Conservation scores are the percentage of sequence identity at the binding interface divided by the percentage of sequence identity outside the interface. Data shown include 30 interfaces for the HM group and 28 interfaces for the HM&HET group (22 homomers and 3 heterodimers of paralogs). P-value is from a Wilcoxon test.

observed that both SSDs and WGDs are more expressed than singletons at the mRNA and protein levels, with WGDs being more expressed than SSDs at the mRNA level (Figure S2A-B). However, expression level (and thus PPI detectability) does not explain completely the enrichment of HMs among duplicated proteins. Both factors, expression and duplication, have significant effects on the probability of proteins to form HMs. It is therefore likely that the overrepresentation of HMs among paralogs is linked to their higher expression along with other factors.

Paralogs that form heteromers tend to have higher sequence identity

The model presented in Figure 1 assumes that the ancestral protein leading to HET formed a HM before duplication. Under the principle of parsimony, we can assume that when at least one paralog forms a HM, the ancestral protein was also a HM. This was shown to be true in general by Diss et al. (2017), who compared yeast WGDs to their orthologs from *Schizosaccharomyces pombe*. To further support this observation, we used PCA to test for HM formation for orthologs from species that diverged prior to the whole-genome duplication event (*Lachancea kluyveri* and *Zygosaccharomyces rouxii*). We looked at paralogs of the mitochondrial translocon complex and the transaldolase, which show HETs according to previous studies (see Materials and methods). We confirm that when one HM was observed in *S. cerevisiae*, at least one ortholog from pre-whole-genome duplication species formed a HM (Figure 2B-C). We also detected interactions between orthologs, suggesting that the ability to interact has been preserved despite the millions of years of evolution separating these species. The absence of interactions for some of these orthologous proteins may be due to the incompatibility of their expression in *S. cerevisiae* or the use of a non-endogenous promoter for these experiments.

We focused on HMs and HETs for 202 pairs of SSDs and 260 pairs of WGDs. This is a reduced dataset compared to the previous section because we needed to consider only pairs for which there was no missing PPI data (see Materials and methods). We combined public data with our own PCA experimental data on 86 SSDs and 149 WGDs (see supplementary text, Figures S3-S4). Overall, the data represents a total of 462 pairs of paralogs (202 SSDs and 260 WGDs) covering 53% of the SSDs and 50% of the WGDs. This dataset encompasses 493 binary interactions of paralogs with themselves (HMs) and 214 interactions with their sister copy (HET).

We classified paralogous pairs into four classes according to whether they show only the HET (HET, 10%), at least one HM but no HET (HM, 39%), at least one of the HM and the HET (HM&HET, 37%) or no interaction (NI, 15%) (Figure 2D, supplementary text). Overall, most pairs forming HETs also form at least one HM (79%). For the rest of the study, we focused our analysis and comparisons on HM and HM&HET pairs because they most likely derive from an ancestral HM. Previous observations showed that paralogs are enriched in protein complexes comprising more than two distinct subunits, partly because these complexes evolved by the initial establishment of self-interactions followed by the duplication of the homomeric proteins (Musso et al., 2007; Pereira-Leal et al., 2007). However, we find that the majority of HM&HET pairs could be simple oligomers of paralogs that do not involve other proteins and are thus not part of large complexes. Only 70 (41%) of the 169 cases of HM&HET are in

complexes with more than two distinct subunits among a set of 5535 complexes reported in databases (see Materials and methods).

We observed that the correlation between HM and HET formation is affected by whether paralogs are SSDs or WGDs (Figure 2E). WGDs tend to form HETs more often when they form at least one HM, resulting in a larger proportion of HM&HET motifs than SSDs. We hypothesize that since SSDs have appeared at different evolutionary times, many of them could be older than WGDs, which could be accompanied by a loss of interactions between paralogs. Indeed, we observed that the distribution of sequence divergence shows lower identity for SSDs than for WGDs, suggesting the presence of ancient duplicates that predate the whole-genome duplication (Figure S5A). Higher protein sequence divergence could lead to the loss of HET complexes because it increases the probability of divergence at the binding interface. We indeed found that among SSDs, those forming HM&HET tend to show a marginally higher overall sequence identity ($p=0.065$, Figure 2F, Figure S5B-C). We also observed a significantly higher sequence identity for WGD pairs forming HM&HET, albeit with a wider distribution (Figure 2F, Figure S5C). This wider distribution derives at least partly from the mixed origin of WGDs (Figure S5D-E). A recent study (Marcet-Houben and Gabaldón, 2015; Wolfe, 2015) showed that WGDs likely have two distinct origins: actual duplication (generating true ohnologs) and hybridization between species (generating homeologs). For pairs whose ancestral state was a HM, we observed that true ohnologs have a tendency to form HET more frequently than homeologs (Figure 2E). Because homeologs had already diverged before the hybridization event, they are older than ohnologs, as shown by their lower pairwise sequence identity (Figure S5D). This observation supports the fact that younger paralogs derived from HMs are more likely to form HETs than older ones.

Amino acid sequence conservation could also have a direct effect on the retention of HETs, independently of the age of the duplication. For instance, among WGDs (either within true ohnologs or homeologs), which all have the same age in their own category, HM&HET pairs have higher sequence identity than HM pairs (Figure S5B, C, E). This is also apparent for pairs of paralogs whose HM or HET structures have been solved by crystallography ($n = 58$ interfaces). Indeed, we found that pairwise amino acid sequence identity was higher for HM&HET than for HM pairs for both entire proteins and for their binding interfaces (Figure 2G). Furthermore, the conservation ratio of the binding interface to the non-interface regions within the available structures is higher for those forming HM&HET, suggesting a causal link between sequence identity at the interface and assembly of HM and HETs (Figure 2H). We extended these analyses to a dataset of human paralogs (Lan and Pritchard, 2016; Singh et al., 2015) to evaluate if these trends can be generalized. Whereas interfaces within PDB structures ($n = 65$ interfaces) are more conserved than the full sequence for both HM and HM&HET motifs (Figure S6A), we did not observe differences in the ratio of conservation of interfaces to non-interfaces (Figure S6B). The reasons for this difference between yeast and humans remain to be explored but it could be caused by mechanisms that do not depend on interfaces to separate paralogous proteins in humans, for instance tissue-specific expression.

Considering that stable interactions are often mediated by protein domains, we looked at the domain composition of paralogs using the Protein Families Database (Pfam) (El-Gebali et al., 2019). We tested

if differences in domain composition could explain the frequency of different interaction motifs. We found that 367 of 448 pairs of paralogs (82%) shared all their domain annotations. Additionally, HM&HET paralogs tend to have more domains in common but the differences are non-significant and appear to be caused by overall sequence divergence (Figure S9A-B). Domain gains and losses are therefore unlikely to contribute to the loss of HET complexes following the duplication of homomers.

Heteromer formation correlates with functional conservation

To test if the retention of HETs correlates with the functional similarity of HM and HM&HET paralogs, we used the similarity of Gene Ontology (GO) terms, reported growth phenotypes of loss-of-function mutants and patterns of genome-wide genetic interactions. These features represent the relationship of genes with cell growth and the gene-gene relationships underlying cell growth. The use of GO terms could bias the analysis because they are often predicted based on sequence features. However, phenotypes and genetic interactions are derived from unbiased experiments because interactions are tested without a priori consideration of a protein's functions (Costanzo et al., 2016). We found that HM&HET pairs are more similar than HM for SSDs (Figure 3 and Figure S10). We observed the same trends for WGDs, although some of the comparisons are either marginally significant or non-significant (Figure 3, comparison between true ohnologs and homeologs in Figure S11). The higher functional similarity observed for HM&HET pairs could be the result of the higher sequence identity described above. However, for a similar level of sequence identity, HM&HET pairs have higher correlation of genetic interaction profiles, higher GO molecular function (for SSDs) and higher GO biological process similarity (for both SSDs and WGDs) than HM pairs (Figure S12). Overall, the retention of HETs after the duplication of HMs appears to correlate with functional similarity, independently from sequence conservation.

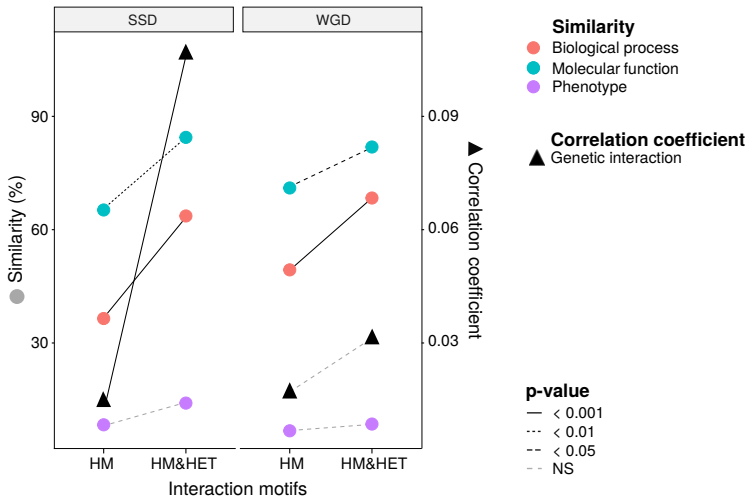

Figure 3: Maintenance of heteromerization between paralogs leads to greater functional similarity. The similarity score is the average proportion of shared terms (100% * Jaccard's index) across pairs of paralogs for GO molecular functions, GO biological processes and gene deletion phenotypes. The mean values of similarity scores and of the correlation of genetic interaction profiles are compared between HM and HM&HET pairs for SSDs and WGDs. P-values are from Wilcoxon tests.

Pleiotropy contributes to the maintenance of heteromers

Since molecular interactions between paralogs predate their functional divergence, it is likely that physical association by itself affects the retention of functional similarity among paralogs. Any feature of paralogs that contributes to the maintenance of the HET state could therefore have a strong impact on the fate of new genes emerging from the duplication of HMs. A large fraction of HMs and HETs use the same binding interface (Bergendahl and Marsh, 2017), so mutations at the interface may have pleiotropic effects on both HMs and HETs (Figure 1), which would lead to correlated responses to selection. If we assume that HMs need to self-interact in order to perform their function, it is expected that natural selection would favor the maintenance of self-assembly. Negative selection on HM interfaces would act on their pleiotropic residues and thus also preserve HET interfaces, preventing the loss of HETs as a correlated response.

We tested this correlated selection model using *in silico* evolution of HM and HET protein complexes (Figure 4A). We used a set of six representative structures of HMs defined as high-quality based on a consensus of different quality measures including the evolutionary conservation of their quaternary structures and thermodynamic models (Dey et al., 2018). We evolved these HM complexes by duplicating them and following the binding energies of the resulting two HMs and HET. We let mutations occur at the binding interface 1) in the absence of selection (neutral model), 2) in the presence of negative selection maintaining only one HM, and 3) with negative selection retaining both HMs. In these three cases, we applied no selection on binding energy of the HET. In the fourth scenario, we applied selection on the HET but not on the HMs to examine if selection maintaining the HET could also favor the retention of HMs. Mutations that have deleterious effects on the complex under selection were lost or allowed to fix with exponentially decaying probability depending on the fitness effect (see Materials and methods) (Figure 4A).

We find that neutral evolution leads to the destabilization of all complexes derived from the simulated duplication of a HM (PDB: 1M38) (Figure 4B), as is expected given that there are more destabilizing mutations than stabilizing ones (Brender and Zhang, 2015; Guerois et al., 2002). Selection to maintain one HM or both HMs significantly slows down the loss of the HET with respect to the neutral scenario (Figure 4C-E). Interestingly, the HET is being destabilized more slowly than the second HM when only one HM is under negative selection. The difficulty of losing the HET in the simulations could explain why for some paralog pairs, only one HM and the HET are preserved, as well as why there are few pairs of paralogs that specifically lose the HET (Figure S13). The reciprocal situation is also true, i.e. negative selection on HET significantly decelerates the loss of stability of both HMs (Figure 4F). These observations hold when simulating the evolution of duplication of five other structures (Figure S14) and when simulating evolution under different combinations of the parameters that control the efficiency of selection and the length of the simulations (Figure S15). By examining the effects that single mutants (only one of the loci gets a nonsynonymous mutation) have on HMs and HET, we find that, as expected, their effects are strongly correlated and thus highly pleiotropic (Pearson's r between 0.64 and 0.9 (Figure S16)). We observe strong pleiotropic effects of mutations for the six structures tested, which explains the correlated responses to selection in the *in silico* evolution.

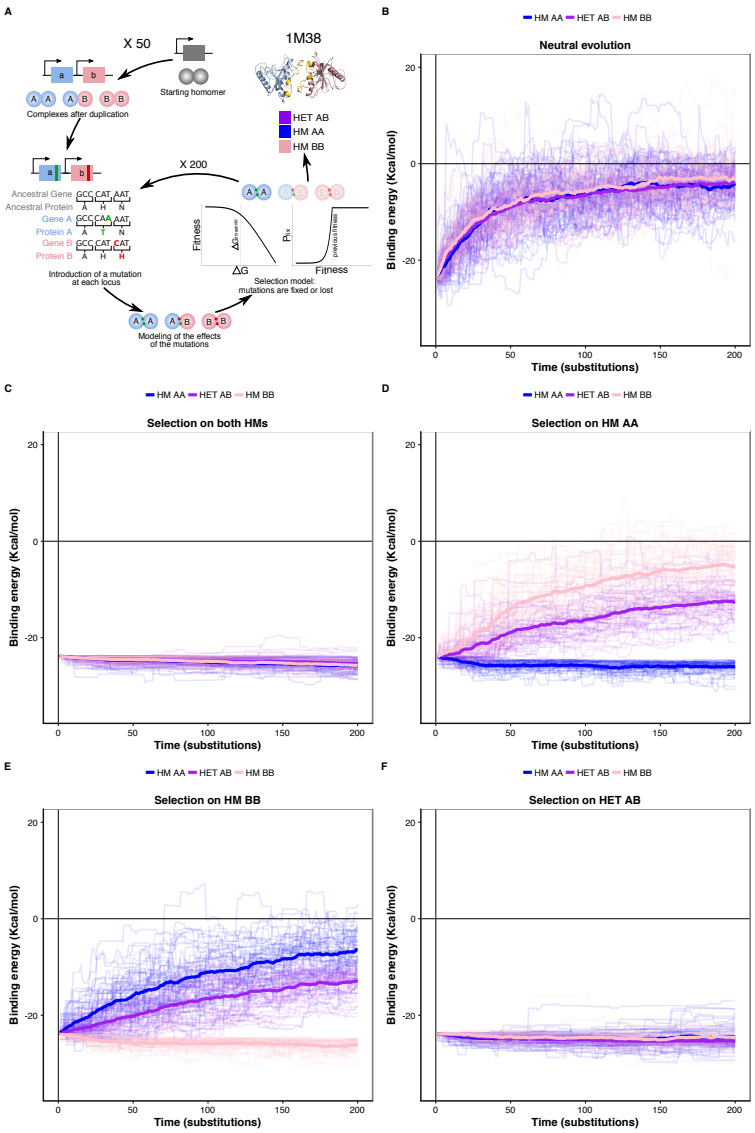

Figure 4: Negative selection to maintain homomers also maintains heteromers.
(A) The duplication of a gene encoding a homomeric protein and the evolution of the complexes is simulated by applying mutations to the corresponding subunits A and B. Only mutations that would require a single nucleotide change are allowed. Stop codons are disallowed. After introducing mutations, the selection model is applied to complexes and mutations are fixed or lost. **(B to F)** The binding energy of the HMs and the HET resulting from the duplication of a HM (PDB: 1M38) is followed through time under different selection regimes applied on protein stability and binding energy. More positive values indicate less favorable binding and more negative values indicate more favorable binding. **(B)** Accumulation and neutral fixation of mutations. **(C)** Selection on both HMs while the HET evolves neutrally. **(D)** Selection on HM AA or **(E)** HM BB: selection maintains one HM while the HET and the other HM evolve neutrally. **(F)** Selection on HET while the HMs evolve neutrally. **(E)** Selection on HM AA or **(F)** HM BB: selection maintains one HM while the HET and the other HM evolve neutrally. Mean binding energies among replicates are shown in thick lines and the individual replicates are shown with thin lines. Fifty replicate populations are monitored in each case and followed for 200 substitutions. PDB structure 1M38 was visualized with PyMOL (Schrödinger LLC, 2015).

Additionally, mutations tend to have greater effects on the HM than on the HET (Figure S16, Figure S17), which agrees with observations on HMs having a greater variance of binding energies than HETs (André et al., 2008; Lukatsky et al., 2007, 2006). As a consequence, HMs that are not under selection in our simulations show higher variability in their binding energy than HETs.

We examined the effects of double mutants (the two loci get a non-synonymous mutation at the interface) on HET formation to study how epistasis may influence the maintenance or loss of HET and HMs when the former or the latter are under selection. We defined epistatic effects as deviations between the observed and the expected effects of mutations on binding energy. Expected effects on HETs were calculated as the average of the effects on the HMs, which each have two subunits with the same mutation. We defined positive epistasis as cases where the observed binding is stronger than expected (more negative $\Delta\Delta G$) and negative epistasis when it is weaker (more positive $\Delta\Delta G$). In terms of evolutionary responses, positive epistasis would contribute to the retention of the HET and

negative epistasis to its loss.

Regardless of the selection scenario, the mutations sampled are slightly enriched for positive epistasis, since the slope values of regression models are smaller than one (0.91 and 0.89 under selection on HMs and HET respectively). When the HMs are maintained by selection, this slightly positive epistasis is also visible in the mutations that are fixed because the epistatic effects are not selected upon. This results in a similar slope for the selected mutations as for the rejected ones. Positive epistasis may therefore contribute to the maintenance of the HET (Figure 5A). On the other hand, selection on the HET results in a further enrichment of mutations with positive epistasis (slope = 0.51, Figure 5B). In this case, mutations tolerated in the HETs and thus fixed are more destabilizing to the HMs. This is also visible in the higher number of fixed substitutions when selection acts on the HET than when it acts on both HMs, particularly for mutations having opposite effects on the HMs (Figure S18). This is also manifested in significantly stronger positive epistasis among fixed pairs of mutations when the HET is under negative selection (t-test, p-value = 0.009). These observations suggest that epistasis may make HETs more robust to mutations than HMs with respect to protein complex assembly, contributing to their maintenance when the HMs are under selection and contributing to the loss of HMs when the HET is under negative selection. This effect is visible in our simulations since selection on the HET results in a slow destabilization of the two HMs (Figure 4, Figure S14), especially when more mutations are attempted (Figure S15), and is observed for all six structures tested (Figure S19).

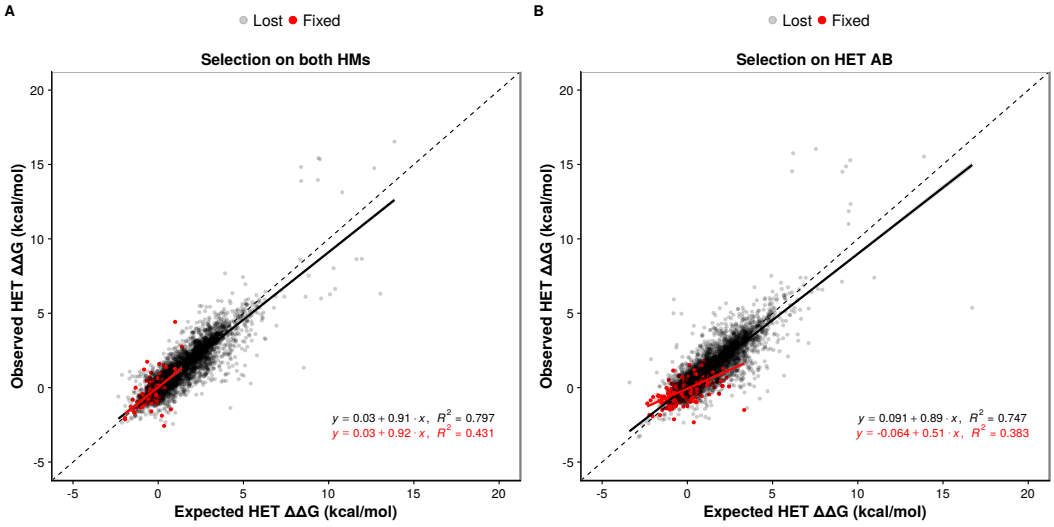

Figure 5: Epistasis favors the maintenance of HETs and the loss of HMs.
(A and B) Observed effects of double mutants on HET (y-axis) are compared to their expected effects (x-axis) based on the average of their effects on the HMs when selection is applied on both HMs (n = 6777 pairs of mutations) **(A)** or on the HET (n = 6760 pairs of mutations) **(B)**. Dashed lines indicate the diagonal for perfect agreement between observations and expectations (no epistasis), black regression lines indicate the best fit for the lost mutants, and red regression lines indicate the best fit for the fixed mutants. Data were obtained from simulations with PDB structure 1M38. The regression coefficients, intercepts and R^2 values are indicated on the figure for fixed and lost mutations. A regression coefficient lower than one means that pairs of mutations have a less destabilizing effects on the HET than expected based on their average effects on the HMs.

Regulatory evolution may break down molecular pleiotropy

The results from simulations show that the loss of HET after the duplication of a HM occurs at a slow rate if HMs are maintained by selection and that specific rare mutations may be required for HETs to be destabilized. However, the simulations only consider the evolution of binding interfaces, which limits the modification of interactions to a subset of all mutations that can ultimately affect PPIs (Hochberg et al., 2018). Other mechanisms that would lead to the loss of HETs could involve transcriptional regulation or cell compartment localization such that paralogs are not present at the same time or in the same cell compartment. To test how regulatory evolution affects interactions, we measured the correlation coefficient of expression profiles of paralogs using mRNA microarray measurements across more than 1000 growth conditions (Ihmels et al., 2004). These expression profiles are more correlated for both SSD and WGD paralogs forming HM&HET than for those forming only HM (p-value = 6.5e-03 and 6.1e-03 respectively, Figure 6A). This result holds using available single-cell RNAseq data (Gasch et al., 2017) although the trend is not significant for WGDs (Figure S20A). Because we found that sequence identity was correlated with both the probability of observing HM&HET and the co-expression of paralogs, we tested if co-expression had an effect on HET formation when controlling for sequence identity. For SSDs, co-expression shows significant effects on HM&HET formation (Figure 6C, Figure S20B) but not for WGDs (Figure 6C, Figure S20B). This is true also when considering the two origins of WGDs separately (Figure S21A-F). The differences of expression correlation between HM and HM&HET could be caused by *cis* regulatory divergence, for instance, HM&HET pairs might have more similar transcription factor binding sites. While we do observe a marginally higher transcription factor binding site similarity for HM&HET pairs than for HM pairs, the tendency is not significant, suggesting other causes for the divergence and similarity of expression profiles (Figure 6B, Figure S22).

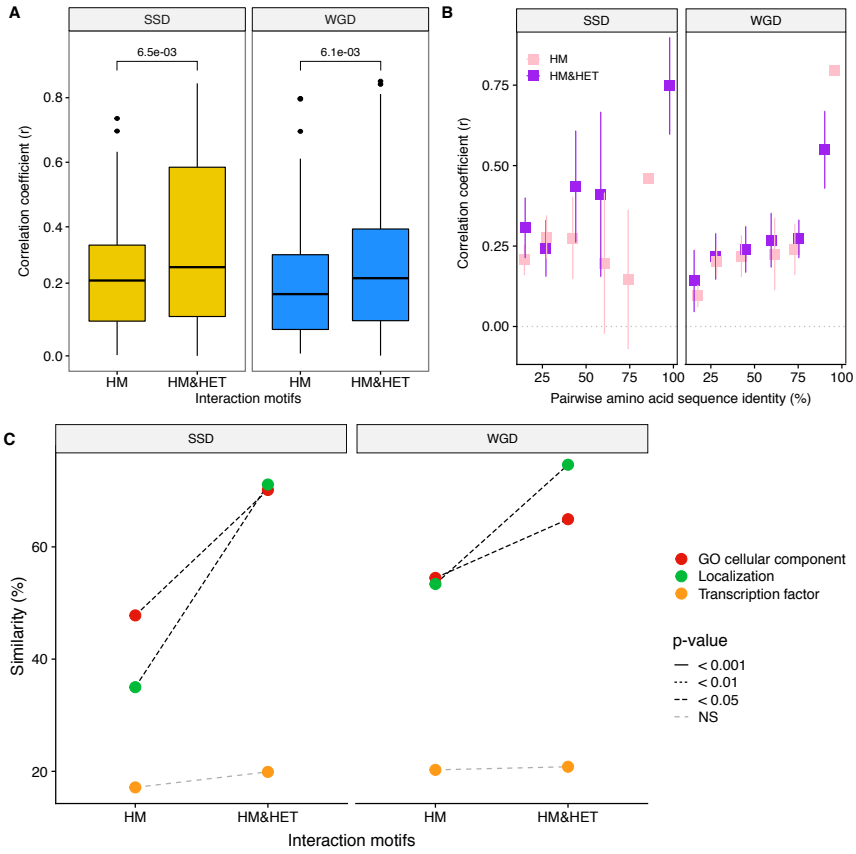

Figure 6: Loss of heteromerization between paralogs may result from regulatory divergence.
(A) Correlation coefficients (Spearman's r) between the expression profiles of paralogs. The data derives from mRNA relative expression across 1000 growth conditions (Ihmels et al., 2004). HM and HM&HET are compared for SSDs (yellow) and WGDs (blue). P-values are from t-tests. **(B)** Correlation of expression profiles between paralogs forming only HM (pink) or HM&HET (purple) as a function of their amino acid sequence identity. The data was binned into six equal categories for representation only. **(C)** Similarity of GO cellular component, GFP-based localization, and transcription factor binding sites (100% * Jaccard's index) are compared between HM and HM&HET for SSDs and WGDs. P-values are from Wilcoxon tests.

Finally, we find that HM&HET paralogs are more similar than HM for both SSDs and WGDs in terms of cellular compartments (GO) and cellular localization derived from experimental data (Figure 6C, Figure S22B-C). For a similar level of sequence identity, HM&HET pairs have higher cellular compartment and cellular localization similarity (for both SSDs and WGDs) than HM pairs (Figure S23). The same tendencies are observed when considering the two classes of WGDs separately (Figure S21G-I).

Overall, coexpression, localization and GO cellular component comparison results suggest that changes in gene and protein regulation could prevent the interaction between paralogs that derive from ancestral HMs, reducing the role of structural pleiotropy in maintaining their associations.

Discussion

Upon duplication, the properties of proteins are inherited from their ancestors, which may affect how paralogs subsequently evolve. Here, we examined the extent to which physical interactions between paralogs are preserved after the duplication of HMs and how these interactions affect functional divergence. Using reported PPI data, crystal structures and new experimental data, we found that paralogs originating from ancestral HMs are more likely to functionally diverge if they lost their ability to form HETs. We propose that non-adaptive mechanisms could play a role in the retention of physical interactions and in turn, impact on functional divergence. By developing a model of *in silico* evolution of PPIs, we found that molecular pleiotropic and epistatic effects of mutations on binding interfaces can constrain the maintenance of HET complexes even if they are not under selection. We hypothesize that this non-adaptive constraint could play a role in slowing down the divergence of paralogs but that it could be counteracted at least partly by regulatory evolution.

The proportions of HMs and HETs among yeast paralogs were first studied more than 15 years ago (Wagner, 2003). It was then suggested that most paralogs forming HETs do not have the ability to form HMs and thus, that evolution of new interactions was rapid. Since then, many PPI experiments have been performed (Chatr-Aryamontri et al., 2017; Kim et al., 2019; Stark et al., 2006; Stynen et al., 2018) and the resulting global picture is different. We found that most of the paralogs forming HETs also form HMs, suggesting that interactions between paralogs are inherited rather than gained *de novo*. This idea is supported by models predicting interaction losses to be much more likely than interaction gains after gene duplication (Gibson and Goldberg, 2009; Presser et al., 2008). Accordingly, the HM&HET state can be more readily achieved by the duplication of an ancestral HM rather than by the duplication of a monomeric protein followed by the gain of the HMs and of the HET. Interacting paralogs are therefore more likely to derive from ancestral HMs, as also shown by Diss et al. (2017) using limited comparative data. For two pairs of *S. cerevisiae* paralogs presenting the HM&HET motif in the literature, we indeed detected HM formation of their orthologs from pre-whole-genome duplication species, supporting the model by which self-interactions and cross-interactions are inherited from the duplication. However, we did not detect HMs for both pre-whole-genome duplication species, which may reflect the incorrect expression of these proteins in *S. cerevisiae* rather than their lack of interaction.

We observed an enrichment of HMs among yeast duplicated proteins compared to singletons, as reported in previous studies (Ispolatov et al., 2005; Pereira-Leal et al., 2007; Pérez-Bercoff et al., 2010; Yang et al., 2003). Also, analyses of PPIs from large-scale experiments have shown that interactions between paralogous proteins are more common than expected by chance (Ispolatov et al., 2005; Musso et al., 2007; Pereira-Leal et al., 2007). Several adaptive hypotheses have been suggested to explain the over-representation of interacting paralogous proteins. For instance, HMs may be preferentially retained, over other duplicates, due to their capacity to provide new adaptive traits by gaining novel functions (neofunctionalization), or by splitting the original ones (subfunctionalization). Similarly, symmetrical HM proteins could have key advantages over monomeric ones for protein stability and

regulation (André et al., 2008; Bergendahl and Marsh, 2017). Levy and Teichmann (2013) suggested that the duplication of HM proteins serves as a seed for the growth of protein complexes. These duplications would allow the diversification of complexes by the asymmetric gain or loss of interactions, which would ultimately lead to the specialization of the duplicates. It is also possible that the presence of HETs itself offers a rapid way to evolve new functions. Examples include bacterial multidrug efflux transporters (Boncoeur et al., 2012) and regulatory mechanisms that evolved this way (Baker et al., 2013; Bridgham et al., 2008; De Smet et al., 2013; Kaltenegger and Ober, 2015). Finally, cotranslational folding has been shown to be a problem for homomeric proteins because of premature assembly of protein complexes, particularly for proteins with interfaces closer to their N-terminus (Natan et al., 2018). The replacement of such HMs by HETs could solve this issue by separating the translation of the proteins to be assembled on two distinct mRNAs.

Non-adaptive mechanisms could also be at play to maintain HETs. Our simulated evolution of the duplication of HMs leads to the proposal of a simple mechanism for the maintenance of HET that does not require adaptive mechanisms. A large fraction of HMs and HETs use the same binding interface (Bergendahl and Marsh, 2017) and as a consequence, negative selection on HM interfaces will also preserve HET interfaces. Our results show that mutations have correlated effects on HM and HET, which slows down the divergence of these complexes. Since some proteins are unstable in the absence of their paralog and lose their capacity to interact with other proteins, cross-stabilization could be another non-adaptive mechanism for the maintenance of the HET (Diss et al., 2017). Notably, these proteins are enriched for paralogs forming HET, suggesting that the individual proteins depend on each other through these physical interactions (Diss et al., 2017). Independent observations by DeLuna et al. (2010) also showed that the deletion of a paralog was sometimes associated with the degradation of the sister copy, particularly among HET paralogs. The Diss et al. and DeLuna et al. observations led to the proposal that paralogs could accumulate complementary degenerative mutations at the structural level after the duplication of a HM (Diss et al., 2017; Kaltenegger and Ober, 2015). This scenario would lead to the maintenance of the HET because destabilizing mutations in one subunit can be compensated by stabilizing mutations in the other, keeping binding energy and overall stability near the optimum. While compensatory mutations could also occur at different positions within identical subunits of the HMs (Uguzzoni et al., 2017), the HET would have access to those same mutations in addition to combinations of mutations in the two paralogous genes. As a result, the number of available compensatory mutations for the HET would be higher than for the HMs.

Furthermore, FoldX in our simulations predicts a slight overall enrichment towards positive epistasis for mutations affecting the two genes whose effects are combined in the HET. This would also contribute to the retention of the HET without adaptive mutations. Together, the smaller effect sizes of individual mutations on HET, the expanded number of compensatory mutations, and the mutational bias toward positive epistasis for the HET observed in our simulations suggest that the assembly of HET might be more robust to mutations than that of HMs. Thus, our simulations show higher potential for the specific retention of the HET than for the specific retention of the two HMs. The next step will be to test these models experimentally.

One of our observations is that WGDs present proportionally more HM&HET motifs than SSDs. We propose that this is at least partly due to the age of paralogs, which would lead to more divergence. This proposal was based on the fact that SSDs in yeast show lower sequence conservation and are thus likely older than WGDs and that even among WGDs, homeologs show the HM&HET motif less frequently than HMs compared to true ohnologs, which are by definition younger. However, the mode of duplication itself could also impact HET maintenance. For instance, upon a whole-genome duplication event, all subunits of complexes are duplicated at the same time, which may contribute to the increased retention of WGDs in complexes compared to SSDs and thus maintain HETs. Indeed, small-scale duplications perturb the stoichiometry of complexes whereas whole-genome duplications preserve it (Birchler and Veitia, 2012; Hakes et al., 2007; Papp et al., 2003; Rice and McLysaght, 2017). In addition, SSDs display higher evolutionary rates than WGDs (Fares et al., 2013), which could lead to the faster loss of their interactions. Another factor that differs is that some WGDs are maintained due to selection for higher gene dosage (Ascencio et al., 2017; Edger and Pires, 2009; Gout and Lynch, 2015; Sugino and Innan, 2006; Thompson et al., 2016). Therefore, the ancestral gene sequence, regulation and function would be conserved, which ultimately favors the maintenance of HETs among WGDs.

We noticed a significant fraction of paralogs forming only HMs but not HET, including some cases of recent duplicates, indicating that the forces maintaining HETs can be overcome. Moreover, although SSDs are more divergent than WGDs on average, sequence divergence and domain composition differ slightly (not significant) between HMs and HM&HETs, suggesting a mechanism other than amino acid sequence divergence for HET loss. Duplicated genes in yeast and other model systems often diverge quickly in terms of transcriptional regulation (Li et al., 2005; Thompson et al., 2013) due to *cis* regulatory mutations (Dong et al., 2011). Because transcriptional divergence of paralogs can directly change PPI profiles, expression changes would be able to rapidly change a motif from HM&HET to HM. Indeed, switching the coding sequences between paralogous loci is sometimes sufficient to change PPI specificity in living cells (Gagnon-Arsenault et al., 2013). Protein localization can also be an important factor affecting the ability of proteins to interact (Rochette et al., 2014). We found that paralogs that derive from HMs and that have lost their ability to form HETs are less co-regulated and less co-localized. This divergence suggests that regulatory evolution could play a role in relieving duplicated homomeric proteins from the correlated effects of mutations affecting shared protein interfaces.

Overall, our analyses show that duplication of self-interacting proteins creates paralogs whose evolution is constrained by pleiotropy in ways that are not expected for monomeric paralogs. Pleiotropy has been known to influence the architecture of complex traits and thus to shape their evolution (Wagner and Zhang, 2011). However, how it takes place at the molecular level and how it can be overcome to allow molecular traits to evolve independently is still largely unknown. Here, we provide a simple system in which the role of pleiotropy can be examined at the molecular level. Because gene duplication is a major mechanism responsible for the evolution of cellular networks and because a large fraction of proteins are oligomeric, the pleiotropic and epistatic constraints described here could be an important force in shaping protein networks. Another important result is that negative selection for the maintenance of heteromers of paralogs is not needed for their preservation, further enhancing our

understanding of the role of non-adaptive evolution in shaping the complexity of cellular structures (Lynch et al., 2014).

Materials and Methods

The protein-protein interactions identified in this publication have been submitted to the IMEx `http://www.imexconsortium.org` consortium through IntAct (Orchard et al., 2014) and are assigned the identifier IM-26944. All scripts used to analyze the data are available at `https://github.com/landrylaboratory/Gene_duplication_2019` (Marchant, 2019; copy archived at `https://github.com/elifesciences-publications/Gene_duplication_2019`).

Characterization of paralogs in *S. cerevisiae* genome

Classification of paralogs by mechanism of duplication

We classified duplicated genes in three categories according to their mechanism of duplication: small-scale duplicates (SSDs); whole-genome duplicates (WGDs) (Byrne and Wolfe, 2005); and doubly duplicated (2D, SSDs and WGDs). We removed WGDs from the paralogs defined in Guan et al. (2007) to generate the list of SSDs. Among paralog pairs with less than 20% of sequence identity in the multiple sequence alignments (Edgar, 2004), we kept only those sharing the same phylome (PhylomeDB; Huerta-Cepas et al., 2008) to make sure they were true paralogs. If one of the two paralogs of an SSD pair was associated to another paralog in a WGD pair, this paralog was considered a 2D. To decrease the potential bias from multiple duplication events, we removed the 2Ds and paralogs from successive small-scale genome duplications from the data on interaction motifs. We used data from Marcet-Houben and Gabaldón (2015) to identify WGDs that are likely true ohnologs or that originated from allopolyploidization (homeologs).

Sequence similarity

Conversion tables between PhylomeDB IDs and systematic yeast IDs were downloaded from `ftp://phylomedb.org/phylomedb/all_id_conversion.txt.gz` on May 15th, 2019. Sequence identity was calculated from multiple sequence alignments from phylome 0003 from PhylomeDB (Huerta-Cepas et al., 2008). The yeast phylome consists of 60 completely sequenced fungal species, with *Homo sapiens* and *Arabidopsis thaliana* as outgroups. Sequences in these phylomes were aligned with MUSCLE v 3.6. When two paralogs were not found in the same multiple sequence alignment from PhylomeDB (32 pairs out of 462 pairs), the sequences were taken from the reference proteome of *S. cerevisiae* assembly R64-1-1 downloaded on April 16th, 2018 from the Ensembl database at `http://useast.ensembl.org/info/data/ftp/index.html` (Zerbino et al., 2018) and realigned to the rest of the phylome with

MUSCLE version 3.8.31 (Edgar, 2004). For six pairs of paralogs that did not have phylomeDB IDs assigned to them, pairwise alignments of their sequences with MUSCLE version 3.8.31 (Edgar, 2004) were used.

Function, transcription factor binding sites, localization of protein complexes, and Pfam annotations

We obtained GO terms (GO slim) from SGD (Cherry et al., 2012) in September 2018. We removed terms corresponding to missing data and created a list of annotations for each SSD and WGD. Annotations were compared to measure the extent of similarity between two members of a pair of duplicates. We calculated the similarity of molecular function, cellular component and biological process taking the number of GO terms in common divided by the total number of unique GO terms of the two paralogs combined (Jaccard index). We compared the same way transcription factor binding sites using YEASTRACT data (Teixeira et al., 2006, 2018), cellular localizations extracted from the Yeast-GFP database (Huh et al., 2003) and many phenotypes associated with the deletion of paralogs (data from SGD in September 2018). For the deletion phenotypes, we kept only information with specific changes (a feature observed and a direction of change relative to wild type). We compared the pairwise correlation of genetic interaction profiles using the genetic interaction profile similarity (measured by Pearson's correlation coefficient) of non-essential genes available in TheCellMap database (version of March 2016) (Usaj et al., 2017). We used the median of correlation coefficients if more than one value was available for a given pair. Non-redundant set of protein complexes was derived from the Complex Portal (Meldal et al., 2015), the CYC2008 catalog (Pu et al., 2007, 2009) and Benschop et al. (2010).

We downloaded Pfam domain annotations (El-Gebali et al., 2019) for the whole *S. cerevisiae* reference proteome on May 2nd, 2019 from the UniprotKB database (Consortium, 2019). We removed pairs of paralogs for which at least one of the proteins had no annotated domains and calculated the Jaccard index.

Homomers and heteromers identified from databases

To complement our experimental data, we extracted HMs and HETs published in BioGRID version BIOGRID-3.5.166 (Chatr-Aryamontri et al., 2012, 2017). We used data derived from the following detection methods: Affinity Capture-MS, Affinity Capture-Western, Reconstituted Complex, Two-hybrid, Biochemical Activity, Co-crystal Structure, Far Western, FRET, Protein-peptide, PCA and Affinity Capture-Luminescence.

It is possible that some HMs or HETs are absent from the database because they have been tested but not detected. This negative information is not reported in databases. We therefore attempted to discriminate non-tested interactions from truly non interacting pairs. A study in which there was not a single HM reported was considered as missing data for all HMs. For both HMs and HETs, the

presence of a protein (or both proteins for HET) as both bait and prey but the absence of interaction was considered as evidence for no interaction. Otherwise, it was considered as missing data.

We also considered data from crystal structures. If a HM was detected in the Protein Data Bank (PDB) (Berman et al., 2000), we inferred that it was present. If the HM was not detected but the monomer was reported, it is likely that there is no HM for this protein and it was thus considered non-HM. If there was no monomer and no HM, the data were considered as missing. We proceeded the same way for HETs.

Data on genome-wide HM screens was obtained from Kim et al. (2019) and Stynen et al. (2018). The two experiments used Protein-fragment complementation assays (PCA), the first one using the dihydrofolate reductase (DHFR) enzyme as a reporter and the second one, a fluorescent protein (also known as Bimolecular fluorescence complementation (BiFC)). We discarded proteins from Stynen et al. (2018) flagged as problematic by Rochette et al. (2014); Stynen et al. (2018); Tarassov et al. (2008) and false positives identified by Kim et al. (2019). All discarded data was considered as missing data. We examined all proteins tested and considered them as HM if they were reported as positive and as non-HM if tested but not reported as positive.

Experimental Protein-fragment complementation assay

We performed a screen using PCA based on DHFR (Tarassov et al., 2008) following standard procedures (Rochette et al., 2015; Tarassov et al., 2008).

DHFR strains

We identified 485 pairs of SSDs and 156 pairs of WGDs present in the Yeast Protein Interactome Collection (Tarassov et al., 2008)and another set of 155 strains constructed by Diss et al. (2017). We retrieved strains from the collection (Tarassov et al., 2008) and we grew them on NAT (DHFR F[1,2] strains) and HygB (DHFR F[3] strains) media. We confirmed the insertion of the DHFR fragments at the correct location by colony PCR using a specific forward Oligo-C targeting a few hundred base pairs upstream of the fusion and a reverse complement oligonucleotide ADHterm_R located in the ADH terminator after the DHFR fragment sequence. Cells from colonies were lysed in 40 µL of 20 mM NaOH for 20 min at 95°C. Tubes were centrifuged for 5 min at 1800 g and 2.5 µL of supernatant was added to a PCR mix composed of 16.85 µL of DNAse free water, 2.5 µL of 10X Taq buffer (BioShop Canada Inc, Canada), 1.5 µL of 25 mM MgCl2, 0.5 µL of 10 mM dNTP (Bio Basic Inc, Canada), 0.15 µL of 5 U/µL Taq DNA polymerase (BioShop Canada Inc, Canada), 0.5 µL of 10 µM Oligo-C and 0.5 µL of 10 µM ADHterm_R. The initial denaturation was performed for 5 min at 95°C and was followed by 35 cycles of 30 s of denaturation at 94°C, 30 s of annealing at 55°C, 1 min of extension at 72°C and by a 3 min final extension at 72°C. We confirmed by PCR 2025 strains from the DHFR collection and 126 strains out of the 154 from Diss et al. (2017).

The missing or non-validated strains were constructed de novo using the standard DHFR strain construction protocol (Michnick et al., 2016; Rochette et al., 2015). The DHFR fragments and associated resistance modules were amplified from plasmids pAG25-linker-F[1,2]-ADHterm (NAT resistance marker) and pAG32-linker-F[3]-ADHterm (HygB resistance marker) (Tarassov et al., 2008) using oligonucleotides. PCR mix was composed of 16.45 µL of DNAse free water, 1 µL of 10 ng/µL plasmid, 5 µL of 5X Kapa Buffer (Kapa Biosystems, Inc, A Roche Company, Canada), 0.75 µL of 10 mM dNTPs, 0.3 µL of 1 U/µL Kapa HiFi HotStart DNA polymerase (Kapa Biosystems, Inc, A Roche Company, Canada) and 0.75 µL of both forward and reverse 10 µM oligos. The initial denaturation was performed for 5 min at 95°C and was followed by 32 cycles of 20 s of denaturation at 98°C, 15 s of annealing at 64.4°C, 2.5 min of extension at 72°C and 5 min of a final extension at 72°C.

We performed strain construction in BY4741 (MAT**a** his3Δ leu2Δ met15Δ ura3Δ) and BY4742 (MATα his3Δ leu2Δ lys2Δ ura3Δ) competent cells prepared as in Gagnon-Arsenault et al. (2013) for the DHFR F[1,2] and DHFR F[3] fusions, respectively. Competent cells (20 µL) were combined with 8 µL of PCR product (0.5–1 µg/µL) and 100 µL of Plate Mixture (PEG3350 40%, 100 mM of LiOAc, 10 mM of Tris-Cl pH 7.5 and 1 mM of EDTA). Cells were vortexed and incubated at room temperature without agitation for 30 min. After adding 15 µL of DMSO and mixing thoroughly, heat shock was performed by incubating in a water bath at 42°C for 15–20 min. Following the heat shock, cells were spun down at 400 g for 3 min. Supernatant was removed by aspiration and cell pellets were resuspended in 100 µL of YPD. Cells were allowed to recover from heat shock for 4 hr at 30°C before being plated on NAT (DHFR F[1,2] strains) or HygB (DHFR F[3] strains) plates. Cells were incubated at 30°C for 3 days. The correct integration of DHFR fragments was confirmed by colony PCR as described above and later by sequencing (Plateforme de séquençage et de génotypage des génomes, CRCHUL, Canada) for specific cases where the interaction patterns suggested a construction problem, for instance when the HET was observed in one direction only or when one HM was missing for a given pair. At the end, we reconstructed and validated 146 new strains. From all available strains, we selected pairs of paralogs for which we had both proteins tagged with both DHFR fragments (four different strains per pair). This resulted in 1172 strains corresponding to 293 pairs of paralogs. We finally discarded pairs considered as leading to false positives by Tarassov et al. (2008), which resulted in 235 pairs.

Construction of DHFR plasmids for orthologous gene expression

For the plasmid-based PCA, Gateway cloning-compatible destination plasmids pDEST-DHFR F[1,2] (TRP1 and LEU2) and pDEST-DHFR F[3] (TRP1 and LEU2) were constructed based on the CEN/ARS low-copy yeast two-hybrid (Y2H) destination plasmids pDEST-AD (TRP1) and pDEST-DB (LEU2) (Rual et al., 2005). A DNA fragment having I-CeuI restriction site was amplified using DEY001 and DEY002 primers without template and another fragment having PI-PspI/I-SceI restriction site was amplified using DEY003 and DEY004 primers without template. pDEST-AD and pDEST-DB plasmids were each digested by PacI and SacI and mixed with the I-CeuI fragment (destined to the PacI locus) and PI-PspI/I-SceI fragment (destined to the SacI locus) for Gibson DNA assembly (Gibson et al., 2009) to generate pDN0501 (TRP1) and pDN0502 (LEU2). Four DNA fragments were then prepared to construct the pDEST-DHFR F[1,2] vectors: (i) a fragment containing the ADH1

promoter; (ii) a fragment containing a Gateway destination site; (iii) a DHFR F[1,2] fragment; and (iv) a backbone plasmid fragment. The ADH1 promoter fragment was amplified from pDN0501 using DEY005 and DEY006 primers and the Gateway destination site fragment was amplified from pDN0501 using DEY007 and DEY008 primers. The DHFR-F[1,2] fragment was amplified from pAG25-linker-F[1,2]-ADHterm (Tarassov et al., 2008) using DEY009 and DEY010 primers.

The backbone fragment was prepared by restriction digestion of pDN0501 or pDN0502 using I-CeuI and PI-PspI and purified by size-selection. The four fragments were assembled by Gibson DNA assembly where each fragment pair was overlapping with more than 30 bp, producing pHMA1001 (TRP1) or pHMA1003 (LEU2). The PstI–SacI region of the plasmids was finally replaced with a DNA fragment containing an amino acid flexible polypeptide linker (GGGGS) prepared by PstI/SacI double digestion of a synthetic DNA fragment DEY011 to produce pDEST-DHFR F[1,2] (TRP1) and pDEST-DHFR F[1,2] (LEU2). The DHFR F[3] fragment was then amplified from pAG32-linker-F[3]-ADHterm with DEY012 and DEY013 primers, digested by SpeI and PI-PspI, and used to replace the SpeI–PI-PspI region of the pDEST-DHFR F[1,2] plasmids, producing pDEST-DHFR F[3] (TRP1) and pDEST-DHFR F[3] (LEU2) plasmids. In this study, we used pDEST-DHFR F[1,2] (TRP1) and pDEST-DHFR F[3] (LEU2) for the plasmid-based DHFR PCA. After Gateway LR cloning of Entry Clones to these destination plasmids, the expression plasmids encode protein fused to the DHFR fragments via an NPAFLYKVVGGGSTS linker.

We obtained the orthologous gene sequences for the mitochondrial translocon complex and the transaldolase proteins of *Lachancea kluyveri* (Kurtzman, 2003) and *Zygosaccharomyces rouxii* (Pribylova et al., 2007) from the Yeast Gene Order Browser (YGOB) (Byrne and Wolfe, 2005). Each ORF was amplified from appropriate gDNA using oligonucleotides. We used 300 ng of purified PCR product to set a BPII recombination reaction (5 µL) into the Gateway Entry Vector pDONR201 (150 ng) according to the manufacturer's instructions (Invitrogen, USA). BPII reaction mix was incubated overnight at 25°C. The reaction was inactivated with proteinase K. The whole reaction was used to transform MC1061 competent *E. coli* cells (Green and Rogers, 2013), followed by selection on solid 2YT medium supplemented with 50 mg/L of kanamycin (BioShop Inc, Canada) at 37°C. Positive clones were detected by PCR using an ORF specific oligonucleotide and a general pDONR201 primer. We then extracted the positive Entry Clones using Presto Mini Plasmid Kit (Geneaid Biotech Ltd, Taiwan) for downstream application.

LRII reactions were performed by mixing 150 ng of the Entry Clone and 150 ng of expression plasmids (pDEST-DHFR F[1,2]-TRP1 or pDEST-DHFR F[3]-LEU2) according to manufacturer's instructions (Invitrogen, USA). The reactions were incubated overnight at 25°C and inactivated with proteinase K. We used the whole reaction to transform MC1061 competent *E. coli* cells, followed by selection on solid 2YT medium supplemented with 100 mg/L ampicillin (BioShop Inc, Canada) at 37°C. Positive clones were confirmed by PCR using a ORF specific primer and a plasmid universal primer. The sequence-verified expression plasmids bearing the orthologous fusions with DHFR F[1,2] and DHFR F[3] fragments were used to transform the yeast strains YY3094 (MAT**a** leu2-3,112 trp1-901 his3-200 ura3-52 gal4Δ gal80Δ LYS2::P_{GAL1}-HIS3 MET2::P_{GAL7}-lacZ cyh2^R can1Δ::P_{CMV}-rtTA-

KanMX4) and YY3095 (MATα leu2-3,112 trp1-901 his3-200 ura3-52 gal4Δ gal80Δ LYS2::P_{GAL1}-HIS3 MET2::PGAL7-lacZ cyh2^R can1Δ::T_{ADH1}-P_{tetO2}-Cre-T_{CYC1}-KanMX4), respectively. Selection was done on SC -trp -ade (YY3094) or on SC -leu -ade (YY3095). The strains YY3094 and YY3095 were generated from BFG-Y2H toolkit strains RY1010 and RY1030 (Yachie et al., 2016), respectively, by restoring their wild type ADE2 genes. The ADE2 gene was restored by homologous recombination of the wild type sequence cassette amplified from the laboratory strain BY4741 using primers DEY014 and DEY015. SC -ade plates were used to obtain successful transformants.

DHFR PCA experiments

Three DHFR PCA experiments were performed, hereafter referred to as PCA1, PCA2 and PCA3. The configuration of strains on plates and the screenings were performed using robotically manipulated pin tools (BM5-SC1, S&P Robotics Inc, Toronto, Canada; Rochette et al., 2015). We first organized haploid strains in 384 colony arrays containing a border of control strains using a cherry-picking 96-pin tool (Figure S7). We constructed four haploid arrays corresponding to paralog 1 and 2 (P1 and P2) and mating type: MAT**a** P1-DHFR F[1,2]; MAT**a** P2-DHFR F[1,2] (on NAT medium); MATα P1-DHFR F[3]; MATα P2-DHFR F[3] (on HygB medium). Border control strains known to show interaction by PCA (MAT**a** LSM8-DHFR F[1-2] and MATα CDC39-DHFR F[3]) were incorporated respectively in all MAT**a** DHFR F[1,2] and MATα DHFR F[3] plates in the first and last columns and rows. The strains were organized as described in Figure S7. The two haploid P1 and P2 384 plates of the same mating type were condensed into a 1536 colony array using a 384-pintool. The two 1536 arrays (one MAT**a** DHFR F[1,2], one MATα DHFR F[3]) were crossed on YPD to systematically test P1-DHFR F[1,2]/P1 DHFR F[3], P1-DHFR F[1,2]/P2-DHFR F[3], P2-DHFR F[1,2]/P1-DHFR F[3] and P2-DHFR F[1,2]/P2-DHFR F[3] interactions in adjacent positions. We performed two rounds of diploid selection (S1 to S2) by replicating the YPD plates onto NAT + HygB and growing for 48 hr. The resulting 1536 diploid plates were replicated twice for 96 hr on DMSO -ade -lys -met control plates (for PCA1 and PCA2) and twice for 96 hr on the selective MTX -ade -lys -met medium (for all runs). Five 1536 PCA plates (PCA1-plate1, PCA1-plate2, PCA2, PCA3-plate1 and PCA3-plate2) were generated this way. We tested the interactions between 277 pairs in five to twenty replicates each.

We also used the robotic platform to generate three bait and three prey 1536 arrays for the DHFR plasmid-based PCA, testing each pairwise interaction at least four times. We mated all MAT**a** DHFR F[1,2] and MATα DHFR F[3] strains on YPD medium at room temperature for 24 hr. We performed two successive steps of diploid selection (SC -leu -trp -ade) followed by two steps on DMSO and MTX media (DMSO -leu -trp -ade and MTX -leu -trp -ade). We incubated the plates of diploid selection at 30°C for 48 hr. Finally, plates from both MTX steps were incubated and monitored for 96 hr at 30°C.

Analysis of DHFR PCA results

Image analysis and colony size quantification

All images were analysed the same way, including images from Stynen et al. (2018). Images of plates were taken with a EOS Rebel T5i camera (Canon, Tokyo, Japan) every two hours during the entire course of the PCA experiments. Incubation and imaging were performed in a spImager custom platform (S&P Robotics Inc, Toronto, Canada). We considered images after two days of growth for diploid selection plates and after four days of growth for DMSO and MTX plates. Images were analysed using gitter (R package version 1.1.1; Wagih and Parts, 2014) to quantify colony sizes by defining a square around the colony center and measuring the foreground pixel intensity minus the background pixel intensity.

Data filtering

For the images from Stynen et al. (2018), we filtered data based on the diploid selection plates. Colonies smaller than 200 pixels were considered as missing data rather than as non-interacting strains. For PCA1, PCA2 and PCA3, colonies flagged as irregular by gitter (as S (colony spill or edge interference) or S, C (low colony circularity) flags) or that did not grow on the last diploid selection step or on DMSO medium (smaller than quantile 25 minus the interquartile range) were considered as missing data. We considered only bait-prey pairs with at least four replicates and used the median of colony sizes as PCA signal. The data was finally filtered based on the completeness of paralogous pairs so we could test HMs and HETs systematically. Thus, we finally obtained results for 241 paralogous pairs. Median colony sizes were log2 transformed after adding a value of 1 to all data to obtain PCA scores. The results of Stynen et al. (2018) and PCA1, PCA2 and PCA3 were strongly correlated (Figure S3B). Similarly, the results correlate well with those reported by Tarassov et al. (2008) (Figure S3C).

Detection of protein-protein interactions

The distribution of PCA scores was modeled per duplication type (SSD and WGD) and per interaction tested (HM or HET) as in Diss et al. (2017) with the *normalmixEM* function (default parameters) available in the R mixtools package (Benaglia et al., 2009). The background signal on MTX was used as a null distribution to which interactions were compared. The size of colonies (PCA scores (PCAs)) were converted to z-scores (Z_s) using the mean (μ_b) and standard deviation (sd_b) of the background distribution (Z_s = (PCA_s - μ_b)/sdb). PPI were considered detected if Zs of the bait-prey pair was greater than 2.5 (Figure S8) (Chrétien et al., 2018).

We observed 24 cases in which only one of the two possible HET interactions was detected (P1-DHFR F[1,2] x P2-DHFR F[3] or P2-DHFR F[1,2] x P1-DHFR F[3]). It is typical for PCA assays to detect interactions in only one orientation or the other (See Tarassov et al. (2008)). However, this could also be caused by one of the four strains having an abnormal fusion sequence. We verified by PCR and sequenced the fusion sequences to make sure this was not the case. The correct strains were conserved and the other ones were re-constructed and retested. No cases of unidirectional HET were observed in our final results. For all 71 pairs after reconstruction, both reciprocal interactions were

detected.

Dataset integration

The PCA data was integrated with other data obtained from databases. The overlaps among the different datasets and the results of our PCA experiments are shown in Figure S4.

Gene expression in MTX condition

Cell cultures for RNAseq

We used the border control diploid strain from the DHFR PCA experiment (MAT**a**/α LSM8-DHFR F[1,2]/LSM8 CDC39/CDC39-DHFR F[3]) to measure expression profile in MTX condition. Three overnight pre-cultures were grown separately in 5 ml of NAT + HygB at 30°C with shaking at 250 rpm. A second set of pre-cultures were grown starting from a dilution at OD600 = 0.01 in 50 ml in the same condition to an OD600 of 0.8 to 1. Final cultures were started at OD600 = 0.03 in 250 ml of synthetic media supplemented with MTX or DMSO (MTX -ade -trp -leu or DMSO -ade -trp -leu) at 30°C with shaking at 250 rpm. These cultures were transferred to 5 × 50 ml tubes when they reached an OD600 of 0.6 to 0.7 and centrifuged at 1008 g at 4°C for 1 min. The supernatant was discarded and cell pellets were frozen in liquid nitrogen and stored at -80°C until processing. RNA extractions and library generation and amplification were performed as described in Eberlein et al. (2019). Briefly, the Quantseq 3' mRNA kit (Lexogen, Vienna, Austria) was used for library preparation (Moll et al., 2014) following the manufacturer's protocol. The PCR cycles number during library amplification was adjusted to 16. The six libraries were pooled and sequenced on a single Ion Torrent chip (ThermoFisher Scientific, Waltham, United States) for a total of 7,784,644 reads on average per library.

RNAseq analysis

Read quality statistics were retrieved from the program FastQC (Andrews, 2010). Reads were cleaned using cutadapt (Martin, 2011). We removed the first 12 bp, trimmed the poly-A tail from the 3' end, trimmed low-quality ends using a cutoff of 15 (phred quality +33) and discarded reads shorter than 30 bp. Raw sequences can be downloaded under the NCBI BioProject ID PRJNA494421.

Cleaned reads were aligned on the reference genome of S288c from SGD (S288C_reference_genome-R64-2-1_20150113.fsa version) using bwa (Li and Durbin, 2009). Because we used a 3'mRNA-Seq Library, reads mapped largely to 3'UTRs. We increased the window of annotated genes in the SGD annotation (saccharomyces_cerevisiae_R64-2-1_20150113.gff version) using the UTR annotation from Nagalakshmi et al. (2008). Based on this reference genes-UTR annotation, the number of mapped reads per genes was estimated using htseq-count of the Python package HTSeq (Anders et al., 2015).

Correlation of gene expression profiles

The correlation of expression profiles for paralogs was calculated using Spearman's correlation from large-scale microarray data (Ihmels et al., 2004) over 1000 mRNA expression profiles from different conditions and different cell cycle phases. These results were compared and confirmed with a large-scale expression data from normalized single-cell RNAseq of S. cerevisiae grown in normal or stressful conditions (0.7 M NaCl) and from different cell cycle phases (Gasch et al., 2017).

Structural analyses

Sequence conservation in binding interfaces of yeast complexes

Identification of crystal structures

The sequences of paralogs classified as SSDs or WGDs (Byrne and Wolfe, 2005; Guan et al., 2007) were taken from the reference proteome of Saccharomyces cerevisiae assembly R64-1-1 and searched using BLASTP (version 2.6.0+) (Camacho et al., 2009) to all the protein sequences contained in the Protein Data Bank (PDB) downloaded on September 21st, 2017 (Berman et al., 2000). Due to the high sequence identity of some paralogs (up to 95%), their structures were assigned as protein subunits from the PDB that had a match with 100% sequence identity and an E-value lower than 1e-6. Only crystal structures that spanned more than 50% of the full protein length were kept for the following analyses. The same method was used to retrieve PDB structures for human paralogous proteins. The human reference proteome Homo_sapiens.GRCh38.pep.all.fa was downloaded on May 16th, 2019 from the Ensembl database (http://useast.ensembl.org/info/data/ftp/index.html) (Zerbino et al., 2018). Pairs of paralogs were retrieved from two different datasets (Lan and Pritchard, 2016; Singh et al., 2015). Protein interactions for those proteins were taken from a merged dataset from the BioGRID (Chatr-Aryamontri et al., 2017) and IntAct (Orchard et al., 2014) databases. The longest protein isoforms for each gene in the dataset were aligned using BLASTP to the set of sequences from the PDB. Matches with 100% sequence identity and E-values below 1e-6 were assigned to the subunits from the PDB structures.

Identification of interfaces

Residue positions involved in protein binding interfaces (contacting and nearby residues) were defined based on the distance of residues to the other subunit (Keskin et al., 2005; Tsai et al., 1996). Contacting residues were defined as those whose two closest non-hydrogen atoms are separated by a distance smaller than the sum of their van der Waals radii plus 0.5 Å. Reference van der Waals radii were obtained with FreeSASA version 2.0.1 (Mitternacht, 2016). Nearby residues are those whose alpha carbons are located at a distance smaller than 6 Å, since it allows inclusion of residues that make up

the scaffold of the interface (Tsai et al., 1996). All distances were measured using the Biopython library (version 1.70) (Cock et al., 2009).

Sequence conservation within interfaces

The dataset of PDB files was filtered to include only the crystallographic structures with the highest resolution available for each complex involving direct contacts between subunits of paralogs. Full-length protein sequences from the reference proteome were then aligned to their matching subunits from the PDB with MUSCLE version 3.8.31 (Edgar, 2004) to assign the structural data to the residues in the full-length protein sequence. These full-length sequences were then aligned to their paralogs and sequences from PhylomeDB (phylome 0003) (Huerta-Cepas et al., 2008) with MUSCLE version 3.8.31. Only three pairs of paralogs that needed realignment were included in this analysis. Sequence identity was calculated within interface regions, which considered the contacting and nearby residues. Paralogs were classified as HM or HM&HET. Pairs of paralogs for which the crystallized domain was only present in one of the proteins were not considered for this analysis.

A similar procedure was applied to the human proteins, with sequences aligned to their corresponding PhylomeDB phylogenies from phylome 0076 resulting from forward and reverse alignments obtained with MUSCLE 3.8, MAFFT v6.712b and DIALIGN-TX, and merged with M-COFFEE (Huerta-Cepas et al., 2008). Considering that human genes code for multiple isoforms, we took the isoforms from the two paralogs that had the highest sequence identity with respect to the PDB structure. When a gene coded for multiple isoforms that were annotated with identical protein sequence in the human reference proteome, we only kept one of them. This resulted in a set of 40 HM interfaces and 25 HM&HET interfaces for a total of 54 different pairs (35 HM pairs and 19 HM&HET). Pairs of paralogs were classified as HM or HM&HET.

Simulations of coevolution of protein complexes

Mutation sampling during evolution of protein interfaces

Simulations were carried out with high-quality crystal structures of homodimeric proteins from PDB (Berman et al., 2000). Four of them (PDB: 1M38, 2JKY, 3D8X, 4FGW) were taken from the above data set of structures that matched yeast paralogs and two others from the same tier of high-quality structures (PDB: 1A82, 2O1V). The simulations model the duplication of the gene encoding the homodimer, giving rise to separate copies that can accumulate different mutations, leading to the formation of HMs and HETs as in Figure 1.

Mutations were introduced using a transition matrix whose substitution probabilities consider the genetic code and allow only substitutions that would require a single base change in the underlying

codons (Thorvaldsen, 2016). Due to the degenerate nature of the genetic code, the model also allows synonymous mutations. Thus, the model explores the effects of amino acid substitutions in both loci, as well as in one locus only. The framework assumes equal mutation rates at both loci, as it proposes a mutation at each locus after every step in the simulation, with 50 replicate populations of 200 steps of substitution in each simulation. Restricting the mutations to the interface maintains sequence identity above 40%, which has been described previously as the threshold at which protein fold remains similar (Addou et al., 2009; Todd et al., 2001; Wilson et al., 2000).

Implementation of selection

Simulations were carried out using the FoldX suite version 4 (Guerois et al., 2002; Schymkowitz et al., 2005). Starting structures were repaired with the RepairPDB function, mutations were simulated with BuildModel followed by the Optimize function, and estimations of protein stability and binding energy of the complex were done with the Stability and Analyse Complex functions, respectively. Effects of mutations on complex fitness were calculated using methods previously described (Kachroo et al., 2015). The fitness of a complex was calculated from three components based on the stability of protein subunits and the binding energy of the complex using Equation 1:

$$x_{\mathrm{i}}^{\mathrm{k}} = -log[e^{\beta*(\Delta \mathrm{G}_{\mathrm{i}}^{\mathrm{k}} - \Delta \mathrm{G}_{\mathrm{threshold}}^{\mathrm{k}})} + 1] \tag{1}$$

where i is the index of the current substitution, k is the index of one of the model's three energetic parameters (stability of subunit A, stability of subunit B, or binding energy of the complex), x_i^k is the fitness component of the kth parameter for the ith substitution, β is a parameter that determines the smoothness of the fitness curve, ΔG_i^k is the free energy value of the k^{th} free energy parameter (stability of subunit A, stability of subunit B, or binding energy of the complex) for the i^{th} substitution, and $\Delta G_{threshold}^k$ is a threshold around which the fitness component starts to decrease. The total fitness of the complex after the i^{th} mutation was calculated as the sum of the three computed values for x_i^k, as shown in Equation 2:

$$x_i = \sum_{n=1}^{3} x_i^k \tag{2}$$

The fitness values of complexes were then used to calculate the probability of fixation (p_{fix}) or rejection of the substitutions using the Metropolis criterion, as in Equation 3:

$$p_{\mathrm{fix}} = \begin{cases} 1 & x_{\mathrm{j}} > x_{\mathrm{i}} \\ e^{-2\mathrm{N}*(\mathrm{x_i} - \mathrm{x_j})} & x_{\mathrm{j}} \leq x_{\mathrm{i}} \end{cases} \tag{3}$$

where p_{fix} is the probability of fixation, x_i is the total fitness value for the complex after i substitutions; x_j is the total fitness value for the complex after j substitutions, with $j = i + 1$; and N is the population size, which influences the efficiency of selection.

Different selection scenarios were examined depending on the complexes whose binding energy and subunit stabilities were under selection: neutral evolution (no selection applied on subunit stability and on the binding energy of the complex), selection on one homodimer, selection on the two homodimers, and selection on the heterodimer. β was set to 10, N was set to 1000 and the $\Delta G^{k}_{threshold}$ were set to 99.9% of the starting values for each complex, following the parameters described in Kachroo et al. (2015). For the simulations with neutral evolution, β was set to 0. For simulations with other combinations of parameters, we varied β and N, one at a time, with β taking values of 1 and 20 and N taking values of 100 and 10000. The simulations with 500 substitutions were carried out with β set to 10, and N set to 1000.

Analyses of simulations

The results from the simulations were then analyzed by distinguishing mutational steps with only one non-synonymous mutation (single mutants, between 29% and 34% of the steps in the simulations) from steps with two non-synonymous mutations (double mutants, between 61% and 68% of the steps). The global data was used to follow the evolution of binding energies of the complexes over time, which are shown in Figure 4. The effects of mutations in HM and HET were compared using the single mutants (Figure S17). The double mutants were used to analyze epistatic and pleiotropic effects (Figure 55, Figure S19) and to compare the rates of mutation fixation based on their effects on the HMs (Figure S18).

Author contributions

CRL, AM and AFC designed this study. AM, AKD, IGA, DA, SA, CE and DEY performed the experiments. AFC performed the *in silico* evolution experiments and the analysis of protein structures. AM, AFC, HAJ and CRL analysed the results. CRL and NY supervised the research. AM, AFC and CRL wrote the manuscript with input from all authors.

Acknowledgements

This work was supported by Canadian Institutes of Health Research grants 299432, 324265 and 387697 to CRL. AM was supported by a FRQS postdoctoral scholarship. AFC was supported by fellowships from PROTEO, MITACS, and Université Laval, as well as joint funding from MEES and AMEXCID. SA was supported by an NSERC undergraduate scholarship. CRL holds the Canada Research Chair in Evolutionary Cells and Systems Biology. We thank SW Michnick for sharing data before publication. The authors thank Philippe Després, Johan Hallin and Anna Fijarczyk for comments on the paper, Rohan Dandage for both comments on the paper and assistance on gathering the data for human paralogs, Rong Shi for useful discussions, and Stéphane Larose for assistance on data management.

Competing interests

The authors have no competing interests to declare.

2 Discussion

Gene duplication has been shown to play a prominent role in the generation of biological innovations. One of the ways in which it can affect protein function is by altering protein-protein interaction networks (He and Zhang, 2005; Pereira-Leal et al., 2007). However, a surprising outcome observed in some yeast proteins is that they become dependent on the presence of their paralogs to be stable and to interact with other proteins (DeLuna et al., 2010; Diss et al., 2017). A potential model that could explain these observations is that the paralogs could have originated from the duplication of a homomeric protein and subsequently evolved to retain only the heteromer. Nevertheless, the factors that could lead to this outcome are not known. Thus, we set out to investigate the extent to which paralogs interact in the yeast protein-protein interaction network by means of experiments, and to simulate the evolution of protein complexes to study how mutations and natural selection can influence the retention of protein complexes. In our simulations, we tested different scenarios to study if the effect of natural selection on the retention of a specific protein complex could influence the retention of the others.

Different models have suggested that the loss of interactions is more frequent than the gain of interactions after gene duplication (Gibson and Goldberg, 2009; Presser et al., 2008). In the light of these results, the duplication of homomeric proteins represents a way to introduce new complexes with different contacts (Pereira-Leal et al., 2007), which can then be lost rather than gained. Given the abundance of homomeric proteins in proteomes (32.48% of the total tested proteins in our dataset), this mechanism is expected to happen often and offers a more parsimonious approach to the generation of diversity and novel complexes. This is confirmed by our observations that pairs of paralogs that form both homomers and heteromers do have self-interacting orthologs in yeast species in which the gene duplications did not occur. We then tested if the proteins in *S. cerevisiae* were capable of interacting with their orthologs in the other yeast species. Remarkably, the proteins are capable of interacting with one another, even though they have evolved separately for millions of years. This result, similar to the complementation by orthologs observed by other studies (Kachroo et al., 2015), suggests that selective pressures acting to maintain the binding between paralogs can also maintain binding to orthologous proteins. This highlights how proteins need to assemble into complexes in order to carry out their functions.

Once the new interactions are established following gene duplication, the two proteins are expected to diverge in sequence. Ashenberg et al. (2011) proposed that a potential mechanism for the development of specificity in protein interactions relies on the previous accumulation of neutral mutations. Indeed, we observed that pairs of paralogs that retain the capability to form heteromers have a higher sequence identity, both at the whole protein level and the interface level. This is consistent with previous observations that proteins have an inherent propensity to self-interact (Lukatsky et al., 2007, 2006). As sequences diverge with time, a trend between the retention of heteromers, sequence conservation, and the age of the duplication was observed, with older paralogs losing the heteromer more often and the newer paralogs retaining it. Our data also reflect the effect of WGD and SSD on the

retention of heteromeric complexes, with duplicates from the WGD being more likely to retain it. Retention of heteromers has important functional consequences, as pairs of paralogs that retain the heteromer are more similar to each other in terms of function, biological processes, phenotypes, and genetic interactions than the ones that do not. A potential outcome of this higher functional similarity could be that the paralogs become dependent on each other to carry out any functions acquired by the heteromer, as in the examples reported by Cui et al. (2015), Baker et al. (2013) and Diss et al. (2017). Remarkably, some of the studied HM&HET paralog pairs lost one of the homomers without losing the heteromer, which suggests that there could be different evolutionary constraints on the complexes.

In order to test potential ways to reach each outcome (HM, HET, HM&HET), we turned to the results of the evolutionary simulations. We observed that the high similarity of the subunits that make up the three protein complexes results in highly correlated evolutionary trajectories. This results in a constraint that slows down the exploration of sequence space of the complexes because they are not independent of each other, particularly when selection is applied either on the heteromer or on both homomers. Nevertheless, when selection is applied on one of the homomers and both the other homomer and the heteromer are considered neutral, we observed that the neutral homomer is destabilized more quickly than the neutral heteromer. This scenario would lead towards the patterns of retained heteromers with the loss of a homomer. Such a scenario could allow the heteromer to evolve for a longer time until it develops a new function or the second gene is lost. Indeed, previous models for the possibility of subfunctionalization and neofunctionalization (Rastogi and Liberles, 2005; Teufel et al., 2016) show that the probability of functional changes increases with the time of retention of the paralogs. In our simulations, we observed that heteromers of paralogs are often preserved throughout, so they could drift neutrally until functions are split or new functions are acquired.

Since the homomers and heteromers in our simulations share subunits, we then looked into the extent of pleiotropy between the complexes. It is characterized by a high correlation of the effects of single mutants on both the homomer and the heteromer, but the magnitude of the effects of mutations was greater for the homomer. This result agrees with previous early remarks that the presence of a mutation on both chains of a homomeric complex should amplify its effects (Monod et al., 1965) and that the distribution of binding energies for homomers has a lower mean (higher stability) but a greater variance than that for heteromers (André et al., 2008; Lukatsky et al., 2007, 2006). While these observations were initially used to explain why homomers are more likely to appear *de novo*, they would also suggest that mutations should have a greater effect on them than on heteromers. In the absence of gene duplication, selective pressure would act on the homomer against the deleterious mutants, preserving the functional state in the population. However, after gene duplication, the initial functional redundancy might reduce the selective pressure on the complexes. Then, as the paralogs accumulate mutations, homomers would be subject to a greater variance of effects. This explains why in our evolutionary simulations the heteromer is preserved under selection for both homomers, favored when under selection for the heteromer, and destabilized more slowly than the neutral homomer under selection for one homomer. Thus, the pleiotropic constraint observed in our simulations could be a major force in the retention of heteromeric complexes.

Nevertheless, a potential way to overcome the pleiotropic constraint would be to separate the expression of the protein chains. As such, we tested for the coexpression of pairs of paralogs. We observed that paralogs that interact with each other are more often coexpressed than those that do not. Multiple sources of information confirm this, including correlations in expression levels across different conditions and similarity in localization component and GO cellular component terms. Furthermore, we observed a higher similarity of transcription factor binding sites for interacting paralogs. The consequences of the regulatory evolution of paralogs have been studied in different models. Some paralog pairs are regulated such that the less expressed copy is slowly reduced until its expression is lost altogether (Gout and Lynch, 2015). Protein-protein interactions are also dependent on regulation, as exchanging the loci of pairs of paralogs can lead them to exchange places in protein complexes (Gagnon-Arsenault et al., 2013). In vertebrates, paralogs can become expressed only in specific tissues, which can stop them from compensating each other's loss. Thus, mutations that affect their function result in tissue-specific diseases (Barshir et al., 2018). Our results show that this mechanism to avoid crosstalk between paralogs is also present in yeast and could contribute to the avoidance of interactions and result in the separation of the paralogs' evolution.

The balance between regulatory evolution and the pleiotropic constraint becomes a determining factor in the retention of heteromers of paralogs. On one hand, the retention of heteromers for longer timescales might contribute to the development of new functions, while the separation of the proteins to prevent interactions might avoid cross-talk and dosage imbalance. Further studies should emphasize on the contributions of each mechanism to cellular robustness, as they can have mixed effects. For example, the pleiotropic constraint might contribute to paralogs being able to buffer each other's loss (DeLuna et al., 2008; Hsiao and Vitkup, 2008), but it might also introduce dependency between paralogs that produces a more deleterious effect upon the loss of one of them (Dandage and Landry, 2019; DeLuna et al., 2010; Diss et al., 2017). Similarly, regulatory separation could be a way to avoid dosage imbalance but could also prevent compensation between paralogs as in tissue-specific diseases (Barshir et al., 2018). Looking at the the different properties of each case of duplicates could help disentangle the complex contributions of gene duplicates to cellular robustness.

A particularly interesting conclusion of this work is that there are intrinsic differences in the effects of mutations for homomers and heteromers. Besides contributing to the retention of protein-protein interactions, mutations having smaller effects on heteromers than on homomers could have implications for the robustness of cellular networks to mutations because smaller effects would be less likely to disrupt complexes. As disruption of protein complexes has been previously linked to diseases (Sahni et al., 2015; Schuster-Böckler and Bateman, 2008), a way to look at this would be to look at the distribution of disease-causing mutations in protein complexes. Schuster-Böckler and Bateman (2008) designed a pipeline to map disease-causing mutations to protein interaction interfaces and identified fewer mutations in heteromeric interfaces than homomeric ones. However, as the PDB is enriched for homomeric complexes, it was not possible to evaluate whether this trend emerged because disease-causing mutations on heteromers could not be mapped since the crystallographic structure had not yet been solved. A way to address this bias would be to look at databases on neutral standing variation (Karczewski et al., 2019). Indeed, preliminary analyses (data not shown) point to disease-causing

mutations being overall enriched in homomeric interfaces, even when controlling for proteins that participate in both homomers and heteromers. However, classifying protein complexes according to other structural variables reveals particular subsets in which it is actually the heteromers that are enriched in disease-causing mutations (data not shown). Further analyses on this topic will yield information on the nature of disease-causing mutations, considerations for protein design, the contribution of heteromers to cellular robustness, and on the evolutionary success of diploid and polyploid genomes for multicellular organisms.

Another perspective of this work would be studying how mutations happen and if there is an overall direction of neutral evolution. In our simulations, most of the sampled mutations were destabilizing for both the homomer and the heteromer. Therefore, most of the available substitutions would inherently lead towards the destabilization of the complexes in the absence of selection. Similar observations were pointed out by Lynch (2018) in the sense that mutational input could have an opposite overall directionality to selection. As such, if selection is not efficient enough, populations might drift away from the overall optimum. A consequence of this model would be that traits in a population stabilize at slightly suboptimal values, which is supported by prokaryotic enzymes having better kinetics than eukaryotic ones, as expected by differences in the effective population sizes (Bar-Even et al., 2011).

A particular bias that was not explored in the simulations but could play a prominent role in this could be mutational bias towards AT. This trend has been shown to be generalized in bacteria, regardless of genomic GC content (Hershberg and Petrov, 2010). Interestingly, even though the genetic code is organized in a way that minimizes changes in hydrophobicity (Freeland and Hurst, 1998; Haig and Hurst, 1991), the AT bias in the genetic code leads to the appearance of more hydrophobic amino acid residues, which would result in higher chain stability but also in higher propensity to interact (Bastolla et al., 2017, 2004). Preliminary analyses based on neutral mutation rates from mutation accumulation experiments in *S. cerevisiae* (Sharp et al., 2018; Zhu et al., 2014) allow us to infer the expected proportions on amino acids in the long term based on neutral mutation rates and the structure of the genetic code. Comparing these long-term expectations to the observed proportions of codons (Nakamura et al., 2000), there is a strong correlation, suggesting that a proteome with a similar amino acid composition to the one observed could be derived solely by neutral evolution and the architecture of the genetic code (data not shown). Looking at the deviations from expectations reveals a strong negative correlation with the propensity of amino acid residues to participate in interfaces, or stickiness, as defined in Levy et al. (2012) (data not shown). In fact, the correlation with stickiness is among the three strongest ones when checking the correlation between the deviation from expectations and 58 other amino acid metrics available in the ExPASy ProtScale server (Gasteiger et al., 2005) (data not shown). This pattern could represent the effect of selection in the proportions of amino acids that make up the proteome as a means to preserve protein structure and function while avoiding the deleterious consequences of the accumulation of hydrophobic residues, such as aggregation. Further studies on this will help identify the role that neutral evolution plays in shaping the genomes and proteomes of different species, as well as the adaptive properties of genomes that have high GC content despite the mutational bias towards AT.

Conclusion

Our work looks at the extent of heteromerization between paralogs in the yeast protein-protein interaction network and discusses mechanisms that can act to preserve or lose these complexes. First of all, we reaffirm the important role of duplications of genes encoding homomeric proteins in the protein-protein interaction network because of the abundance of homomeric proteins and the way these duplications contribute to the generation of new complexes. Then, we characterize the mechanisms that could break up the relationship between the two resulting paralogs, that is: the accumulation of mutations and regulatory evolution. We find that the similarity of the complexes results in mutations having pleiotropic effects on them that constrain their evolution because mutations affect more than one complex. Simulations show that complexes remain stable even when selection does not act on them because of this pleiotropic constraint, with heteromers being particularly able to explore more of the sequence space while remaining stable because mutations tend to have smaller effects on them than on homomers. On the other hand, regulatory evolution could separate the paralogs to make them evolve independently from each other and overcome this pleiotropic constraint. We show that the interplay between these two mechanisms determines the retention of heteromers of paralogs after gene duplication. A consequence of the retention of heteromers can be the change in the proteins' function, which then can have distinct effects on robustness at the molecular and organism level.

Bibliography

Abrusán, G. and Marsh, J. A. (2018). Ligand binding site structure influences the evolution of protein complex function and topology. *Cell Rep.*, 22(12):3265–3276.

Addou, S., Rentzsch, R., Lee, D., and Orengo, C. A. (2009). Domain-based and family-specific sequence identity thresholds increase the levels of reliable protein function transfer. *J. Mol. Biol.*, 387(2):416–430.

Aguzzi, A. and O'Connor, T. (2010). Protein aggregation diseases: pathogenicity and therapeutic perspectives. *Nat. Rev. Drug Discov.*, 9(3):237–248.

Ahnert, S. E., Marsh, J. A., Hernández, H., Robinson, C. V., and Teichmann, S. A. (2015). Principles of assembly reveal a periodic table of protein complexes. *Science*, 350(6266):aaa2245.

Alford, R. F., Leaver-Fay, A., Jeliazkov, J. R., O'Meara, M. J., DiMaio, F. P., Park, H., Shapovalov, M. V., Renfrew, P. D., Mulligan, V. K., Kappel, K., Labonte, J. W., Pacella, M. S., Bonneau, R., Bradley, P., Dunbrack, Jr, R. L., Das, R., Baker, D., Kuhlman, B., Kortemme, T., and Gray, J. J. (2017). The rosetta All-Atom energy function for macromolecular modeling and design. *J. Chem. Theory Comput.*, 13(6):3031–3048.

Amoutzias, G. D., Robertson, D. L., Van de Peer, Y., and Oliver, S. G. (2008). Choose your partners: dimerization in eukaryotic transcription factors. *Trends Biochem. Sci.*, 33(5):220–229.

Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq–a python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169.

André, I., Strauss, C. E. M., Kaplan, D. B., Bradley, P., and Baker, D. (2008). Emergence of symmetry in homooligomeric biological assemblies. *Proc. Natl. Acad. Sci. U. S. A.*, 105(42):16148–16152.

Andrews, S. (2010). FastQC a quality control tool for high throughput sequence data. `http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`.

Arnold, M. L. and Martin, N. H. (2010). Hybrid fitness across time and habitats. *Trends Ecol. Evol.*, 25(9):530–536.

Ascencio, D., Ochoa, S., Delaye, L., and DeLuna, A. (2017). Increased rates of protein evolution and asymmetric deceleration after the whole-genome duplication in yeasts. *BMC Evol. Biol.*, 17(1).

Ashenberg, O., Rozen-Gagnon, K., Laub, M. T., and Keating, A. E. (2011). Determinants of homodimerization specificity in histidine kinases. *J. Mol. Biol.*, 413(1):222–235.

Aury, J.-M., Jaillon, O., Duret, L., Noel, B., Jubin, C., Porcel, B. M., Ségurens, B., Daubin, V., Anthouard, V., Aiach, N., Arnaiz, O., Billaut, A., Beisson, J., Blanc, I., Bouhouche, K., Câmara, F., Duharcourt, S., Guigo, R., Gogendeau, D., Katinka, M., Keller, A.-M., Kissmehl, R., Klotz, C., Koll, F., Le Mouël, A., Lepère, G., Malinsky, S., Nowacki, M., Nowak, J. K., Plattner, H., Poulain, J., Ruiz, F., Serrano, V., Zagulski, M., Dessen, P., Bétermier, M., Weissenbach, J., Scarpelli, C., Schächter, V., Sperling, L., Meyer, E., Cohen, J., and Wincker, P. (2006). Global trends of whole-genome duplications revealed by the ciliate paramecium tetraurelia. *Nature*, 444(7116):171–178.

Baker, C. R., Hanson-Smith, V., and Johnson, A. D. (2013). Following gene duplication, paralog interference constrains transcriptional circuit evolution. *Science*, 342(6154):104–108.

Bar-Even, A., Noor, E., Savir, Y., Liebermeister, W., Davidi, D., Tawfik, D. S., and Milo, R. (2011). The moderately efficient enzyme: evolutionary and physicochemical trends shaping enzyme parameters. *Biochemistry*, 50(21):4402–4410.

Barshir, R., Hekselman, I., Shemesh, N., Sharon, M., Novack, L., and Yeger-Lotem, E. (2018). Role of duplicate genes in determining the tissue-selectivity of hereditary diseases. *PLoS Genet.*, 14(5):e1007327.

Bastolla, U., Dehouck, Y., and Echave, J. (2017). What evolution tells us about protein physics, and protein physics tells us about evolution. *Curr. Opin. Struct. Biol.*, 42:59–66.

Bastolla, U., Moya, A., Viguera, E., and van Ham, R. C. H. J. (2004). Genomic determinants of protein folding thermodynamics in prokaryotic organisms. *J. Mol. Biol.*, 343(5):1451–1466.

Benaglia, T., Chauveau, D., Hunter, D. R., and Young, D. S. (2009). mixtools: an R package for analyzing mixture models. *J. Stat. Softw.*, 32(06).

Benschop, J. J., Brabers, N., van Leenen, D., Bakker, L. V., van Deutekom, H. W. M., van Berkum, N. L., Apweiler, E., Lijnzaad, P., Holstege, F. C. P., and Kemmeren, P. (2010). A consensus of core protein complex compositions for saccharomyces cerevisiae. *Mol. Cell*, 38(6):916–928.

Bergendahl, L. T. and Marsh, J. A. (2017). Functional determinants of protein assembly into homomeric complexes. *Sci. Rep.*, 7(1):4932.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The protein data bank. *Nucleic Acids Res.*, 28(1):235–242.

Birchler, J. A. and Veitia, R. A. (2012). Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. *Proc. Natl. Acad. Sci. U. S. A.*, 109(37):14746–14753.

Boeckmann, B., Blatter, M.-C., Famiglietti, L., Hinz, U., Lane, L., Roechert, B., and Bairoch, A. (2005). Protein variety and functional diversity: Swiss-Prot annotation in its biological context. *C. R. Biol.*, 328(10-11):882–899.

Bomblies, K. and Madlung, A. (2014). Polyploidy in the arabidopsis genus. *Chromosome Res.*, 22(2):117–134.

Boncoeur, E., Durmort, C., Bernay, B., Ebel, C., Di Guilmi, A. M., Croizé, J., Vernet, T., and Jault, J.-M. (2012). PatA and PatB form a functional heterodimeric ABC multidrug efflux transporter responsible for the resistance of streptococcus pneumoniae to fluoroquinolones. *Biochemistry*, 51(39):7755–7765.

Braasch, I., Gehrke, A. R., Smith, J. J., Kawasaki, K., Manousaki, T., Pasquier, J., Amores, A., Desvignes, T., Batzel, P., Catchen, J., Berlin, A. M., Campbell, M. S., Barrell, D., Martin, K. J., Mulley, J. F., Ravi, V., Lee, A. P., Nakamura, T., Chalopin, D., Fan, S., Wcisel, D., Cañestro, C., Sydes, J., Beaudry, F. E. G., Sun, Y., Hertel, J., Beam, M. J., Fasold, M., Ishiyama, M., Johnson, J., Kehr, S., Lara, M., Letaw, J. H., Litman, G. W., Litman, R. T., Mikami, M., Ota, T., Saha, N. R., Williams, L., Stadler, P. F., Wang, H., Taylor, J. S., Fontenot, Q., Ferrara, A., Searle, S. M. J., Aken, B., Yandell, M., Schneider, I., Yoder, J. A., Volff, J.-N., Meyer, A., Amemiya, C. T., Venkatesh, B., Holland, P. W. H., Guiguen, Y., Bobe, J., Shubin, N. H., Di Palma, F., Alföldi, J., Lindblad-Toh, K., and Postlethwait, J. H. (2016). The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. *Nat. Genet.*, 48(4):427–437.

Brender, J. R. and Zhang, Y. (2015). Predicting the effect of mutations on Protein-Protein binding interactions through Structure-Based interface profiles. *PLoS Comput. Biol.*, 11(10):e1004494.

Brenner, S., Jacob, F., and Meselson, M. (1961). An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature*, 190:576–581.

Bridgham, J. T., Brown, J. E., Rodríguez-Marí, A., Catchen, J. M., and Thornton, J. W. (2008). Evolution of a new function by degenerative mutation in cephalochordate steroid receptors. *PLoS Genet.*, 4(9):e1000191.

Byrne, K. P. and Wolfe, K. H. (2005). The yeast gene order browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res.*, 15(10):1456–1461.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, 10:421.

Celaj, A., Schlecht, U., Smith, J. D., Xu, W., Suresh, S., Miranda, M., Aparicio, A. M., Proctor, M., Davis, R. W., Roth, F. P., and St Onge, R. P. (2017). Quantitative analysis of protein interaction network dynamics in yeast. *Mol. Syst. Biol.*, 13(7):934.

Charron, G., Marsit, S., Henault, M., Martin, H., and Landry, C. (2019). Spontaneous whole-genome duplication restores fertility in interspecific hybrids. *Nature Communications*, 10:4126.

Chatr-Aryamontri, A., Breitkreutz, B.-J., Heinicke, S., Boucher, L., Winter, A., Stark, C., Nixon, J., Ramage, L., Kolas, N., O'Donnell, L., and Others (2012). The BioGRID interaction database: 2013 update. *Nucleic Acids Res.*, 41(D1):D816–D823.

Chatr-Aryamontri, A., Oughtred, R., Boucher, L., Rust, J., Chang, C., Kolas, N. K., O'Donnell, L., Oster, S., Theesfeld, C., Sellam, A., Stark, C., Breitkreutz, B.-J., Dolinski, K., and Tyers, M. (2017). The BioGRID interaction database: 2017 update. *Nucleic Acids Res.*, 45(D1):D369–D379.

Cherry, J. M., Hong, E. L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E. T., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S. R., Fisk, D. G., Hirschman, J. E., Hitz, B. C., Karra, K., Krieger, C. J., Miyasato, S. R., Nash, R. S., Park, J., Skrzypek, M. S., Simison, M., Weng, S., and Wong, E. D. (2012). Saccharomyces genome database: the genomics resource of budding yeast. *Nucleic Acids Res.*, 40(Database issue):D700–5.

Chrétien, A.-È., Gagnon-Arsenault, I., Dubé, A. K., Barbeau, X., Després, P. C., Lamothe, C., Dion-Côté, A.-M., Lagüe, P., and Landry, C. R. (2018). Extended linkers improve the detection of protein-protein interactions (PPIs) by dihydrofolate reductase protein-fragment complementation assay (DHFR PCA) in living cells. *Mol. Cell. Proteomics*, 17(2):373–383.

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M. J. L. (2009). Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423.

Consortium, U. (2019). Uniprot: a worldwide hub of protein knowledge. *Nucleic acids research*, 47(D1):D506–D515.

Costa, T. R. D., Ignatiou, A., and Orlova, E. V. (2017). Structural analysis of protein complexes by cryo electron microscopy. *Methods Mol. Biol.*, 1615:377–413.

Costanzo, M., VanderSluis, B., Koch, E. N., Baryshnikova, A., Pons, C., Tan, G., Wang, W., Usaj, M., Hanchard, J., Lee, S. D., Pelechano, V., Styles, E. B., Billmann, M., van Leeuwen, J., van Dyk, N., Lin, Z.-Y., Kuzmin, E., Nelson, J., Piotrowski, J. S., Srikumar, T., Bahr, S., Chen, Y., Deshpande, R., Kurat, C. F., Li, S. C., Li, Z., Usaj, M. M., Okada, H., Pascoe, N., San Luis, B.-J., Sharifpoor, S., Shuteriqi, E., Simpkins, S. W., Snider, J., Suresh, H. G., Tan, Y., Zhu, H., Malod-Dognin, N., Janjic, V., Przulj, N., Troyanskaya, O. G., Stagljar, I., Xia, T., Ohya, Y., Gingras, A.-C., Raught, B., Boutros, M., Steinmetz, L. M., Moore, C. L., Rosebrock, A. P., Caudy, A. A., Myers, C. L., Andrews, B., and Boone, C. (2016). A global genetic interaction network maps a wiring diagram of cellular function. *Science*, 353(6306).

Crow, J. F. and Kimura, M. (1970). *An introduction to population genetics theory.* New York, Evanston and London: Harper & Row, Publishers.

Cui, J., Xiao, J., Tagliabracci, V. S., Wen, J., Rahdar, M., and Dixon, J. E. (2015). A secretory kinase complex regulates extracellular protein phosphorylation. *eLife*, 4:e06120.

Dandage, R. and Landry, C. R. (2019). Paralog dependency indirectly affects the robustness of human cells. *Molecular systems biology*, 15(9).

De Smet, R., Adams, K. L., Vandepoele, K., Van Montagu, M. C. E., Maere, S., and Van de Peer, Y. (2013). Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proc. Natl. Acad. Sci. U. S. A.*, 110(8):2898–2903.

Dehal, P. and Boore, J. L. (2005). Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biology*, 3(10):e314.

DeLuna, A., Springer, M., Kirschner, M. W., and Kishony, R. (2010). Need-based up-regulation of protein levels in response to deletion of their duplicate genes. *PLoS Biol.*, 8(3):e1000347.

DeLuna, A., Vetsigian, K., Shoresh, N., Hegreness, M., Colón-González, M., Chao, S., and Kishony, R. (2008). Exposing the fitness contribution of duplicated genes. *Nat. Genet.*, 40(5):676–681.

DePristo, M. A., Weinreich, D. M., and Hartl, D. L. (2005). Missense meanderings in sequence space: a biophysical view of protein evolution. *Nat. Rev. Genet.*, 6(9):678–687.

Dey, S., Ritchie, D. W., and Levy, E. D. (2018). PDB-wide identification of biological assemblies from conserved quaternary structure geometry. *Nat. Methods*, 15(1):67–72.

Diss, G., Gagnon-Arsenault, I., Dion-Coté, A.-M., Vignaud, H., Ascencio, D. I., Berger, C. M., and Landry, C. R. (2017). Gene duplication can impart fragility, not robustness, in the yeast protein interaction network. *Science*, 355(6325):630–634.

Diss, G. and Lehner, B. (2018). The genetic landscape of a physical interaction. *eLife*, 7:e32472.

Dong, D., Yuan, Z., and Zhang, Z. (2011). Evidences for increased expression variation of duplicate genes in budding yeast: from cis- to trans-regulation effects. *Nucleic Acids Res.*, 39(3):837–847.

Dufton, M. J. (1983). The significance of redundancy in the genetic code. *J. Theor. Biol.*, 102(4):521–526.

Eberlein, C., Hénault, M., Fijarczyk, A., Charron, G., Bouvier, M., Kohn, L. M., Anderson, J. B., and Landry, C. R. (2019). Hybridization is a recurrent evolutionary stimulus in wild yeast speciation. *Nat. Commun.*, 10(1):923.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, 32(5):1792–1797.

Edger, P. P. and Pires, J. C. (2009). Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. *Chromosome Res.*, 17(5):699–717.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., et al. (2019). The pfam protein families database in 2019. *Nucleic acids research*, 47(D1):D427–D432.

Englander, S. W. and Mayne, L. (2014). The nature of protein folding pathways. *Proc. Natl. Acad. Sci. U. S. A.*, 111(45):15873–15880.

Fares, M. A., Keane, O. M., Toft, C., Carretero-Paulet, L., and Jones, G. W. (2013). The roles of whole-genome and small-scale duplications in the functional specialization of saccharomyces cerevisiae genes. *PLoS genetics*, 9(1):e1003176.

Ferrada, E. and Wagner, A. (2008). Protein robustness promotes evolutionary innovations on large evolutionary time-scales. *Proc. Biol. Sci.*, 275(1643):1595–1602.

Firnberg, E. and Ostermeier, M. (2013). The genetic code constrains yet facilitates darwinian evolution. *Nucleic Acids Res.*, 41(15):7420–7428.

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. L., and Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151(4):1531–1545.

Freeland, S. J. and Hurst, L. D. (1998). The genetic code is one in a million. *J. Mol. Evol.*, 47(3):238–248.

Freschi, L., Torres-Quiroz, F., Dubé, A. K., and Landry, C. R. (2013). qPCA: a scalable assay to measure the perturbation of protein–protein interactions in living cells. *Mol. Biosyst.*, 9(1):36–43.

Gagnon-Arsenault, I., Marois Blanchet, F.-C., Rochette, S., Diss, G., Dubé, A. K., and Landry, C. R. (2013). Transcriptional divergence plays a role in the rewiring of protein interaction networks after gene duplication. *J. Proteomics*, 81:112–125.

Garcia-Seisdedos, H., Empereur-Mot, C., Elad, N., and Levy, E. D. (2017). Proteins evolve on the edge of supramolecular self-assembly. *Nature*, 548(7666):244–247.

Gasch, A. P., Yu, F. B., Hose, J., Escalante, L. E., Place, M., Bacher, R., Kanbar, J., Ciobanu, D., Sandor, L., Grigoriev, I. V., et al. (2017). Single-cell rna sequencing reveals intrinsic and extrinsic regulatory heterogeneity in yeast responding to stress. *PLoS biology*, 15(12):e2004050.

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D., and Bairoch, A. (2005). Protein identification and analysis tools on the ExPASy server. In Walker, J. M., editor, *The Proteomics Protocols Handbook*, pages 571–607. Humana Press, Totowa, NJ.

Gibson, D. G., Young, L., Chuang, R.-Y., Venter, J. C., Hutchison, 3rd, C. A., and Smith, H. O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods*, 6(5):343–345.

Gibson, T. A. and Goldberg, D. S. (2009). Questioning the ubiquity of neofunctionalization. *PLoS Comput. Biol.*, 5(1):e1000252.

Gout, J.-F., Kahn, D., Duret, L., and Paramecium Post-Genomics Consortium (2010). The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. *PLoS Genet.*, 6(5):e1000944.

Gout, J.-F. and Lynch, M. (2015). Maintenance and loss of duplicated genes by dosage subfunctionalization. *Mol. Biol. Evol.*, 32(8):2141–2148.

Green, R. and Rogers, E. J. (2013). Transformation of chemically competent e. coli. *Methods Enzymol*, 529:329–336.

Gros, F., Hiatt, H., Gilbert, W., Kurland, C. G., Risebrough, R. W., and Watson, J. D. (1961). Unstable ribonucleic acid revealed by pulse labelling of escherichia coli. *Nature*, 190:581–585.

Guan, Y., Dunham, M. J., and Troyanskaya, O. G. (2007). Functional analysis of gene duplications in saccharomyces cerevisiae. *Genetics*, 175(2):933–943.

Guerois, R., Nielsen, J. E., and Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.*, 320(2):369–387.

Haig, D. and Hurst, L. D. (1991). A quantitative measure of error minimization in the genetic code. *J. Mol. Evol.*, 33(5):412–417.

Hakes, L., Pinney, J. W., Lovell, S. C., Oliver, S. G., and Robertson, D. L. (2007). All duplicates are not equal: the difference between small-scale and genome duplication. *Genome Biol.*, 8(10):R209.

He, X. and Zhang, J. (2005). Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics*, 169(2):1157–1164.

Hershberg, R. and Petrov, D. A. (2010). Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genet.*, 6(9):e1001115.

Hochberg, G. K. A., Shepherd, D. A., Marklund, E. G., Santhanagoplan, I., Degiacomi, M. T., Laganowsky, A., Allison, T. M., Basha, E., Marty, M. T., Galpin, M. R., Struwe, W. B., Baldwin, A. J., Vierling, E., and Benesch, J. L. P. (2018). Structural principles that enable oligomeric small heat-shock protein paralogs to evolve distinct functions. *Science*, 359(6378):930–935.

Hollingsworth, S. A. and Dror, R. O. (2018). Molecular dynamics simulation for all. *Neuron*, 99(6):1129–1143.

Hsiao, T.-L. and Vitkup, D. (2008). Role of duplicate genes in robustness against deleterious human mutations. *PLoS Genet.*, 4(3):e1000014.

Huerta-Cepas, J., Bueno, A., Dopazo, J., and Gabaldón, T. (2008). PhylomeDB: a database for genome-wide collections of gene phylogenies. *Nucleic Acids Res.*, 36(Database issue):D491–6.

Huh, W.-K., Falvo, J. V., Gerke, L. C., Carroll, A. S., Howson, R. W., Weissman, J. S., and O'Shea, E. K. (2003). Global analysis of protein localization in budding yeast. *Nature*, 425(6959):686–691.

Ihmels, J., Bergmann, S., and Barkai, N. (2004). Defining transcription modules using large-scale gene expression data. *Bioinformatics*, 20(13):1993–2003.

Innan, H. and Kondrashov, F. (2010). The evolution of gene duplications: classifying and distinguishing between models. *Nat. Rev. Genet.*, 11(2):97–108.

Ispolatov, I., Yuryev, A., Mazo, I., and Maslov, S. (2005). Binding properties and evolution of homodimers in protein–protein interaction networks. *Nucleic Acids Res.*, 33(11):3629–3635.

Jacob, F. and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.*, 3:318–356.

Janin, J., Bahadur, R. P., and Chakrabarti, P. (2008). Protein–protein interaction and quaternary structure. *Q. Rev. Biophys.*, 41(2):133–180.

Kachroo, A. H., Laurent, J. M., Yellman, C. M., Meyer, A. G., Wilke, C. O., and Marcotte, E. M. (2015). Systematic humanization of yeast genes reveals conserved functions and genetic modularity. *Science*, 348(6237):921–925.

Kaltenegger, E. and Ober, D. (2015). Paralogue interference affects the dynamics after gene duplication. *Trends Plant Sci.*, 20(12):814–821.

Kaniecki, K., De Tullio, L., and Greene, E. C. (2018). A change of view: homologous recombination at single-molecule resolution. *Nat. Rev. Genet.*, 19(4):191–207.

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., Seaby, E. G., Kosmicki, J. A., Walters, R. K., Tashman, K., Farjoun, Y., Banks, E., Poterba, T., Wang, A., Seed, C., Whiffin, N., Chong, J. X., Samocha, K. E., Pierce-Hoffman, E., Zappala, Z., O'Donnell-Luria, A. H., Minikel, E. V., Weisburd, B., Lek, M., Ware, J. S., Vittal, C., Armean, I. M., Bergelson, L., Cibulskis, K., Connolly, K. M., Covarrubias, M., Donnelly, S., Ferriera, S., Gabriel, S., Gentry, J., Gupta, N., Jeandet, T., Kaplan, D., Llanwarne, C., Munshi, R., Novod, S., Petrillo, N., Roazen, D., Ruano-Rubio, V., Saltzman, A., Schleicher, M., Soto, J., Tibbetts, K., Tolonen, C., Wade, G., Talkowski, M. E., The Genome Aggregation Database Consortium, Neale, B. M., Daly, M. J., and MacArthur, D. G. (2019). Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv*, page 531210.

Kellis, M., Birren, B. W., and Lander, E. S. (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast saccharomyces cerevisiae. *Nature*, 428(6983):617–624.

Keskin, O., Ma, B., and Nussinov, R. (2005). Hot regions in protein–protein interactions: the organization and contribution of structurally conserved hot spot residues. *Journal of molecular biology*, 345(5):1281–1294.

Kim, Y., Jung, J. P., Pack, C.-G., and Huh, W.-K. (2019). Global analysis of protein homomerization in saccharomyces cerevisiae. *Genome Res.*, 29(1):135–145.

Kleckner, I. R. and Foster, M. P. (2011). An introduction to NMR-based approaches for measuring protein dynamics. *Biochim. Biophys. Acta*, 1814(8):942–968.

Korostelev, A., Trakhanov, S., Laurberg, M., and Noller, H. F. (2006). Crystal structure of a 70S ribosome-tRNA complex reveals functional interactions and rearrangements. *Cell*, 126(6):1065–1077.

Kurtzman, C. P. (2003). Phylogenetic circumscription of saccharomyces, kluyveromyces and other members of the saccharomycetaceae, and the proposal of the new genera lachancea, nakaseomyces, naumovia, vanderwaltozyma and zygotorulaspora. *FEMS yeast research*, 4(3):233–245.

Lan, X. and Pritchard, J. K. (2016). Coregulation of tandem duplicate genes slows evolution of subfunctionalization in mammals. *Science*, 352(6288):1009–1013.

Landry, C. R., Levy, E. D., Abd Rabbo, D., Tarassov, K., and Michnick, S. W. (2013). Extracting insight from noisy cellular networks. *Cell*, 155(5):983–989.

Lev, I., Volpe, M., and Ben-Aroya, S. (2014). Identification of genes important for the physical interaction between protein pairs through reverse pca (rpca). *Current protocols in cell biology*, 64(1):1–11.

Levy, E. D., Boeri Erba, E., Robinson, C. V., and Teichmann, S. A. (2008). Assembly reflects evolution of protein complexes. *Nature*, 453(7199):1262–1265.

Levy, E. D., De, S., and Teichmann, S. A. (2012). Cellular crowding imposes global constraints on the chemistry and evolution of proteomes. *Proc. Natl. Acad. Sci. U. S. A.*, 109(50):20461–20466.

Levy, E. D. and Teichmann, S. (2013). Structural, evolutionary, and assembly principles of protein oligomerization. In Giraldo, J. and Ciruela, F., editors, *Progress in Molecular Biology and Translational Science*, volume 117, pages 25–51. Academic Press.

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., and Pfister, H. (2014). UpSet: Visualization of intersecting sets. *IEEE Trans. Vis. Comput. Graph.*, 20(12):1983–1992.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760.

Li, W.-H., Yang, J., and Gu, X. (2005). Expression divergence between duplicate genes. *Trends Genet.*, 21(11):602–607.

Liberles, D. A., Teichmann, S. A., Bahar, I., Bastolla, U., Bloom, J., Bornberg-Bauer, E., Colwell, L. J., de Koning, A. P. J., Dokholyan, N. V., Echave, J., Elofsson, A., Gerloff, D. L., Goldstein, R. A., Grahnen, J. A., Holder, M. T., Lakner, C., Lartillot, N., Lovell, S. C., Naylor, G., Perica, T., Pollock, D. D., Pupko, T., Regan, L., Roger, A., Rubinstein, N., Shakhnovich, E., Sjölander, K., Sunyaev, S., Teufel, A. I., Thorne, J. L., Thornton, J. W., Weinreich, D. M., and Whelan, S. (2012). The interface of protein structure, protein biophysics, and molecular evolution. *Protein Sci.*, 21(6):769–785.

Lien, S., Koop, B. F., Sandve, S. R., Miller, J. R., Kent, M. P., Nome, T., Hvidsten, T. R., Leong, J. S., Minkley, D. R., Zimin, A., Grammes, F., Grove, H., Gjuvsland, A., Walenz, B., Hermansen, R. A., von Schalburg, K., Rondeau, E. B., Di Genova, A., Samy, J. K. A., Olav Vik, J., Vigeland, M. D., Caler, L., Grimholt, U., Jentoft, S., Våge, D. I., de Jong, P., Moen, T., Baranski, M., Palti, Y., Smith, D. R., Yorke, J. A., Nederbragt, A. J., Tooming-Klunderud, A., Jakobsen, K. S., Jiang,

X., Fan, D., Hu, Y., Liberles, D. A., Vidal, R., Iturra, P., Jones, S. J. M., Jonassen, I., Maass, A., Omholt, S. W., and Davidson, W. S. (2016). The atlantic salmon genome provides insights into rediploidization. *Nature*, 533(7602):200–205.

Lin, M., Whitmire, S., Chen, J., Farrel, A., Shi, X., and Guo, J.-T. (2017). Effects of short indels on protein structure and function in human genomes. *Sci. Rep.*, 7(1):9313.

Lukatsky, D. B., Shakhnovich, B. E., Mintseris, J., and Shakhnovich, E. I. (2007). Structural similarity enhances interaction propensity of proteins. *J. Mol. Biol.*, 365(5):1596–1606.

Lukatsky, D. B., Zeldovich, K. B., and Shakhnovich, E. I. (2006). Statistically enhanced self-attraction of random patterns. *Phys. Rev. Lett.*, 97(17):178101.

Lynch, M. (2012). The evolution of multimeric protein assemblages. *Mol. Biol. Evol.*, 29(5):1353–1366.

Lynch, M. (2018). Phylogenetic divergence of cell biological features. *eLife*, 7.

Lynch, M., Field, M. C., Goodson, H. V., Malik, H. S., Pereira-Leal, J. B., Roos, D. S., Turkewitz, A. P., and Sazer, S. (2014). Evolutionary cell biology: two origins, one objective. *Proc. Natl. Acad. Sci. U. S. A.*, 111(48):16990–16994.

Lynch, M. and Force, A. (2000). The probability of duplicate gene preservation by subfunctionalization. *Genetics*, 154(1):459–473.

Maity, H., Maity, M., Krishna, M. M. G., Mayne, L., and Englander, S. W. (2005). Protein folding: the stepwise assembly of foldon units. *Proc. Natl. Acad. Sci. U. S. A.*, 102(13):4741–4746.

Marcet-Houben, M. and Gabaldón, T. (2015). Beyond the Whole-Genome duplication: Phylogenetic evidence for an ancient interspecies hybridization in the baker's yeast lineage. *PLoS Biol.*, 13(8):e1002220.

Marsh, J. A. and Teichmann, S. A. (2014). Protein flexibility facilitates quaternary structure assembly and evolution. *PLoS Biol.*, 12(5):e1001870.

Marsh, J. A. and Teichmann, S. A. (2015). Structure, dynamics, assembly, and evolution of protein complexes. *Annu. Rev. Biochem.*, 84:551–575.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10–12.

Mayer, M. P. (2010). Gymnastics of molecular chaperones. *Mol. Cell*, 39(3):321–331.

Meldal, B. H. M., Forner-Martinez, O., Costanzo, M. C., Dana, J., Demeter, J., Dumousseau, M., Dwight, S. S., Gaulton, A., Licata, L., Melidoni, A. N., Ricard-Blum, S., Roechert, B., Skyzypek, M. S., Tiwari, M., Velankar, S., Wong, E. D., Hermjakob, H., and Orchard, S. (2015). The complex portal–an encyclopaedia of macromolecular complexes. *Nucleic Acids Res.*, 43(Database issue):D479–84.

Michnick, S. W., Levy, E. D., Landry, C. R., Kowarzyk, J., and Messier, V. (2016). The dihydrofolate reductase Protein-Fragment complementation assay: A Survival-Selection assay for Large-Scale analysis of Protein–Protein interactions. *Cold Spring Harb. Protoc.*, 2016(11):db.prot090027.

Mitternacht, S. (2016). FreeSASA: An open source C library for solvent accessible surface area calculations. *F1000Res.*, 5:189.

Miyazawa, S. and Jernigan, R. L. (1996). Residue – residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.*, 256(3):623–644.

Moll, P., Ante, M., Seitz, A., and Reda, T. (2014). QuantSeq 3' mRNA sequencing for RNA quantification. *Nat. Methods*, 11(12):i–iii.

Monod, J., Wyman, J., and Changeux, J. P. (1965). On the nature of allosteric transitions: a plausible model. *J. Mol. Biol.*, 12:88–118.

Musso, G., Zhang, Z., and Emili, A. (2007). Retention of protein complex membership by ancient duplicated gene products in budding yeast. *Trends Genet.*, 23(6):266–269.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 320(5881):1344–1349.

Nakamura, Y., Gojobori, T., and Ikemura, T. (2000). Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Res.*, 28(1):292.

Natan, E., Endoh, T., Haim-Vilmovsky, L., Flock, T., Chalancon, G., Hopper, J. T. S., Kintses, B., Horvath, P., Daruka, L., Fekete, G., Pál, C., Papp, B., Oszi, E., Magyar, Z., Marsh, J. A., Elcock, A. H., Babu, M. M., Robinson, C. V., Sugimoto, N., and Teichmann, S. A. (2018). Cotranslational protein assembly imposes evolutionary constraints on homomeric proteins. *Nat. Struct. Mol. Biol.*, 25(3):279–288.

Ohno, S. (1970). Duplication for the sake of producing more of the same. In *Evolution by Gene Duplication*, pages 59–65. Springer.

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N. H., Chavali, G., Chen, C., del Toro, N., Duesbury, M., Dumousseau, M., Galeota, E., Hinz, U., Iannuccelli, M., Jagannathan, S., Jimenez, R., Khadake, J., Lagreid, A., Licata, L., Lovering, R. C., Meldal, B., Melidoni, A. N., Milagros, M., Peluso, D., Perfetto, L., Porras, P., Raghunath, A., Ricard-Blum, S., Roechert, B., Stutz, A., Tognolli, M., van Roey, K., Cesareni, G., and Hermjakob, H. (2014). The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.*, 42(D1):D358–D363.

Orr, H. A. (2009). Fitness and its role in evolutionary genetics. *Nature Reviews Genetics*, 10(8):531.

Otto, S. P. and Whitton, J. (2000). Polyploid incidence and evolution. *Annu. Rev. Genet.*, 34:401–437.

Pandey, A. V., Henderson, C. J., Ishii, Y., Kranendonk, M., Backes, W. L., and Zanger, U. M. (2017). Editorial: Role of Protein-Protein interactions in metabolism: Genetics, structure, function. *Frontiers in Pharmacology*, 8.

Papp, B., Pál, C., and Hurst, L. D. (2003). Dosage sensitivity and the evolution of gene families in yeast. *Nature*, 424(6945):194–197.

Parmley, J. L. and Hurst, L. D. (2007). How do synonymous mutations affect fitness? *Bioessays*, 29(6):515–519.

Payne, J. L. and Wagner, A. (2019). The causes of evolvability and their evolution. *Nat. Rev. Genet.*, 20(1):24–38.

Pereira-Leal, J. B., Levy, E. D., Kamp, C., and Teichmann, S. A. (2007). Evolution of protein complexes by duplication of homomeric interactions. *Genome Biol.*, 8(4):R51.

Pérez-Bercoff, A., Makino, T., and McLysaght, A. (2010). Duplicability of self-interacting human genes. *BMC Evol. Biol.*, 10:160.

Presser, A., Elowitz, M. B., Kellis, M., and Kishony, R. (2008). The evolutionary dynamics of the saccharomyces cerevisiae protein interaction network after duplication. *Proc. Natl. Acad. Sci. U. S. A.*, 105(3):950–954.

Pribylova, L., de Montigny, J., and Sychrova, H. (2007). Osmoresistant yeast zygosaccharomyces rouxii: the two most studied wild-type strains (atcc 2623 and atcc 42981) differ in osmotolerance and glycerol metabolism. *Yeast*, 24(3):171–180.

Pu, S., Vlasblom, J., Emili, A., Greenblatt, J., and Wodak, S. J. (2007). Identifying functional modules in the physical interactome of saccharomyces cerevisiae. *Proteomics*, 7(6):944–960.

Pu, S., Wong, J., Turner, B., Cho, E., and Wodak, S. J. (2009). Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Res.*, 37(3):825–831.

Rastogi, S. and Liberles, D. A. (2005). Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *BMC Evol. Biol.*, 5:28.

Reams, A. B. and Roth, J. R. (2015). Mechanisms of gene duplication and amplification. *Cold Spring Harb. Perspect. Biol.*, 7(2):a016592.

Rice, A. M. and McLysaght, A. (2017). Dosage-sensitive genes in evolution and disease. *BMC Biol.*, 15(1):78.

Rochette, S., Diss, G., Filteau, M., Leducq, J.-B., Dubé, A. K., and Landry, C. R. (2015). Genome-wide protein-protein interaction screening by protein-fragment complementation assay (PCA) in living cells. *J. Vis. Exp.*, (97).

Rochette, S., Gagnon-Arsenault, I., Diss, G., and Landry, C. R. (2014). Modulation of the yeast protein interactome in response to DNA damage. *J. Proteomics*, 100:25–36.

Rouquier, S., Blancher, A., and Giorgi, D. (2000). The olfactory receptor gene repertoire in primates and mouse: evidence for reduction of the functional fraction in primates. *Proc. Natl. Acad. Sci. U. S. A.*, 97(6):2870–2874.

Rual, J.-F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G. F., Gibbons, F. D., Dreze, M., Ayivi-Guedehoussou, N., Klitgord, N., Simon, C., Boxem, M., Milstein, S., Rosenberg, J., Goldberg, D. S., Zhang, L. V., Wong, S. L., Franklin, G., Li, S., Albala, J. S., Lim, J., Fraughton, C., Llamosas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski, R. S., Vandenhaute, J., Zoghbi, H. Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette-Stamm, L., Cusick, M. E., Hill, D. E., Roth, F. P., and Vidal, M. (2005). Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062):1173–1178.

Sahni, N., Yi, S., Taipale, M., Fuxman Bass, J. I., Coulombe-Huntington, J., Yang, F., Peng, J., Weile, J., Karras, G. I., Wang, Y., Kovács, I. A., Kamburov, A., Krykbaeva, I., Lam, M. H., Tucker, G., Khurana, V., Sharma, A., Liu, Y.-Y., Yachie, N., Zhong, Q., Shen, Y., Palagi, A., San-Miguel, A., Fan, C., Balcha, D., Dricot, A., Jordan, D. M., Walsh, J. M., Shah, A. A., Yang, X., Stoyanova, A. K., Leighton, A., Calderwood, M. A., Jacob, Y., Cusick, M. E., Salehi-Ashtiani, K., Whitesell, L. J., Sunyaev, S., Berger, B., Barabási, A.-L., Charloteaux, B., Hill, D. E., Hao, T., Roth, F. P., Xia, Y., Walhout, A. J. M., Lindquist, S., and Vidal, M. (2015). Widespread macromolecular interaction perturbations in human genetic disorders. *Cell*, 161(3):647–660.

Sandve, S. R., Rohlfs, R. V., and Hvidsten, T. R. (2018). Subfunctionalization versus neofunctionalization after whole-genome duplication. *Nat. Genet.*, 50(7):908–909.

Schlecht, U., Liu, Z., Blundell, J. R., Onge, R. P. S., and Levy, S. F. (2017). A scalable double-barcode sequencing platform for characterization of dynamic protein-protein interactions. *Nature communications*, 8:15586.

Schuster-Böckler, B. and Bateman, A. (2008). Protein interactions in human genetic diseases. *Genome Biol.*, 9(1):R9.

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., and Serrano, L. (2005). The FoldX web server: an online force field. *Nucleic Acids Research*, 33(Web Server):W382–W388.

Scott, J. D. and Pawson, T. (2009). Cell signaling in space and time: where proteins come together and when they're apart. *Science*, 326(5957):1220–1224.

Sharp, N. P., Sandell, L., James, C. G., and Otto, S. P. (2018). The genome-wide rate and spectrum of spontaneous mutations differ between haploid and diploid yeast. *Proc. Natl. Acad. Sci. U. S. A.*, 115(22):E5046–E5055.

Shi, Y. (2014). A glimpse of structural biology through x-ray crystallography. *Cell*, 159(5):995–1014.

Singh, P. P., Arora, J., and Isambert, H. (2015). Identification of ohnolog genes originating from whole genome duplication in early vertebrates, based on synteny comparison across multiple genomes. *PLoS computational biology*, 11(7):e1004394.

Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, 34(Database issue):D535–9.

Starr, T. N. and Thornton, J. W. (2016). Epistasis in protein evolution. *Protein Science*, 25(7):1204–1218.

Stynen, B., Abd-Rabbo, D., Kowarzyk, J., Miller-Fleming, L., Aulakh, S. K., Garneau, P., Ralser, M., and Michnick, S. W. (2018). Changes of cell biochemical states are revealed in protein homomeric complex dynamics. *Cell*, 175(5):1418–1429.e9.

Sugino, R. P. and Innan, H. (2006). Selection for more of the same product as a force to enhance concerted evolution of duplicated genes. *Trends Genet.*, 22(12):642–644.

Tarassov, K., Messier, V., Landry, C. R., Radinovic, S., Serna Molina, M. M., Shames, I., Malitskaya, Y., Vogel, J., Bussey, H., and Michnick, S. W. (2008). An in vivo map of the yeast protein interactome. *Science*, 320(5882):1465–1470.

Teixeira, M. C., Monteiro, P., Jain, P., Tenreiro, S., Fernandes, A. R., Mira, N. P., Alenquer, M., Freitas, A. T., Oliveira, A. L., and Sá-Correia, I. (2006). The YEASTRACT database: a tool for the analysis of transcription regulatory associations in saccharomyces cerevisiae. *Nucleic Acids Res.*, 34(Database issue):D446–51.

Teixeira, M. C., Monteiro, P. T., Palma, M., Costa, C., Godinho, C. P., Pais, P., Cavalheiro, M., Antunes, M., Lemos, A., Pedreira, T., and Sá-Correia, I. (2018). YEASTRACT: an upgraded database for the analysis of transcription regulatory networks in saccharomyces cerevisiae. *Nucleic Acids Research*, 46(D1):D348–D353.

Teufel, A. I., Johnson, M. M., Laurent, J. M., Kachroo, A. H., Marcotte, E. M., and Wilke, C. O. (2019). The many nuanced evolutionary consequences of duplicated genes. *Mol. Biol. Evol.*, 36(2):304–314.

Teufel, A. I., Liu, L., and Liberles, D. A. (2016). Models for gene duplication when dosage balance works as a transition state to subsequent neo-or sub-functionalization. *BMC Evol. Biol.*, 16:45.

Thompson, A., Zakon, H. H., and Kirkpatrick, M. (2016). Compensatory drift and the evolutionary dynamics of Dosage-Sensitive duplicate genes. *Genetics*, 202(2):765–774.

Thompson, D. A., Roy, S., Chan, M., Styczynski, M. P., Pfiffner, J., French, C., Socha, A., Thielke, A., Napolitano, S., Muller, P., Kellis, M., Konieczka, J. H., Wapinski, I., and Regev, A. (2013). Correction: Evolutionary principles of modular gene regulation in yeasts. *eLife*, 2.

Thorvaldsen, S. (2016). A mutation model from first principles of the genetic code. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 13(5):878–886.

Titeca, K., Lemmens, I., Tavernier, J., and Eyckerman, S. (2019). Discovering cellular protein-protein interactions: Technological strategies and opportunities. *Mass spectrometry reviews*, 38(1):79–111.

Todd, A. E., Orengo, C. A., and Thornton, J. M. (2001). Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.*, 307(4):1113–1143.

Tsai, C. J., Lin, S. L., Wolfson, H. J., and Nussinov, R. (1996). Protein-protein interfaces: architectures and interactions in protein-protein interfaces and in protein cores. their similarities and differences. *Crit. Rev. Biochem. Mol. Biol.*, 31(2):127–152.

Uguzzoni, G., Lovis, S. J., Oteri, F., Schug, A., Szurmant, H., and Weigt, M. (2017). Large-scale identification of coevolution signals across homo-oligomeric protein interfaces by direct coupling analysis. *Proceedings of the National Academy of Sciences*, 114(13):E2662–E2671.

Usaj, M., Tan, Y., Wang, W., VanderSluis, B., Zou, A., Myers, C. L., Costanzo, M., Andrews, B., and Boone, C. (2017). TheCellMap.org: A Web-Accessible database for visualizing and mining the global yeast genetic interaction network. *G3: Genes|Genomes|Genetics*, 7(5):1539–1549.

Vidal, M., Cusick, M. E., and Barabási, A.-L. (2011). Interactome networks and human disease. *Cell*, 144(6):986–998.

Wagner, A. (2003). How the global structure of protein interaction networks evolves. *Proc. Biol. Sci.*, 270(1514):457–466.

Wagner, G. P. and Zhang, J. (2011). The pleiotropic structure of the genotype–phenotype map: the evolvability of complex organisms. *Nature Reviews Genetics*, 12(3):204–213.

Wan, C., Borgeson, B., Phanse, S., Tu, F., Drew, K., Clark, G., Xiong, X., Kagan, O., Kwan, J., Bezginov, A., Chessman, K., Pal, S., Cromar, G., Papoulas, O., Ni, Z., Boutz, D. R., Stoilova, S., Havugimana, P. C., Guo, X., Malty, R. H., Sarov, M., Greenblatt, J., Babu, M., Derry, W. B., Tillier, E. R., Wallingford, J. B., Parkinson, J., Marcotte, E. M., and Emili, A. (2015). Panorama of ancient metazoan macromolecular complexes. *Nature*, 525(7569):339–344.

Wang, M., Weiss, M., Simonovic, M., Haertinger, G., Schrimpf, S. P., Hengartner, M. O., and von Mering, C. (2012). PaxDb, a database of protein abundance averages across all three domains of life. *Mol. Cell. Proteomics*, 11(8):492–500.

Wilson, C. A., Kreychman, J., and Gerstein, M. (2000). Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *Journal of Molecular Biology*, 297(1):233–249.

Wlodawer, A., Dauter, Z., Porebski, P. J., Minor, W., Stanfield, R., Jaskolski, M., Pozharski, E., Weichenberger, C. X., and Rupp, B. (2018). Detect, correct, retract: How to manage incorrect structural models. *FEBS J.*, 285(3):444–466.

Wolfe, K. H. (2015). Origin of the yeast Whole-Genome duplication. *PLoS Biol.*, 13(8):e1002221.

Xing, J., Wang, H., Belancio, V. P., Cordaux, R., Deininger, P. L., and Batzer, M. A. (2006). Emergence of primate genes by retrotransposon-mediated sequence transduction. *Proc. Natl. Acad. Sci. U. S. A.*, 103(47):17608–17613.

Yachie, N., Petsalaki, E., Mellor, J. C., Weile, J., Jacob, Y., Verby, M., Ozturk, S. B., Li, S., Cote, A. G., Mosca, R., Knapp, J. J., Ko, M., Yu, A., Gebbia, M., Sahni, N., Yi, S., Tyagi, T., Sheykhkarimli, D., Roth, J. F., Wong, C., Musa, L., Snider, J., C. Liu, Y., Yu, H., Braun, P., Stagljar, I., Hao, T., Calderwood, M. A., Pelletier, L., Aloy, P., Hill, D. E., Vidal, M., and Roth, F. P. (2016). Pooled-matrix protein interaction screens using barcode fusion genetics. *Molecular Systems Biology*, 12(4):863–863.

Yang, J., Lusk, R., and Li, W.-H. (2003). Organismal complexity, protein complexity, and gene duplicability. *Proc. Natl. Acad. Sci. U. S. A.*, 100(26):15661–15665.

Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Girón, C. G., Gil, L., Gordon, L., Haggerty, L., Haskell, E., Hourlier, T., Izuogu, O. G., Janacek, S. H., Juettemann, T., To, J. K., Laird, M. R., Lavidas, I., Liu, Z., Loveland, J. E., Maurel, T., McLaren, W., Moore, B., Mudge, J., Murphy, D. N., Newman, V., Nuhn, M., Ogeh, D., Ong,

C. K., Parker, A., Patricio, M., Riat, H. S., Schuilenburg, H., Sheppard, D., Sparrow, H., Taylor, K., Thormann, A., Vullo, A., Walts, B., Zadissa, A., Frankish, A., Hunt, S. E., Kostadima, M., Langridge, N., Martin, F. J., Muffato, M., Perry, E., Ruffier, M., Staines, D. M., Trevanion, S. J., Aken, B. L., Cunningham, F., Yates, A., and Flicek, P. (2018). Ensembl 2018. *Nucleic Acids Res.*, 46(D1):D754–D761.

Zhang, J. (2003). Evolution by gene duplication: an update. *Trends Ecol. Evol.*, 18(6):292–298.

Zhu, Y. O., Siegal, M. L., Hall, D. W., and Petrov, D. A. (2014). Precise estimates of mutation rate and spectrum in yeast. *Proc. Natl. Acad. Sci. U. S. A.*, 111(22):E2310–8.

Annex

Supplementary text

Comparison of PCA results with previous studies

We performed a screen using PCA based on the DHFR (Tarassov et al., 2008) to test for PPIs among SSDs and WGDs in *Saccharomyces cerevisiae*, specifically testing for self-interactions (HMs) and interactions between paralogs (HETs). The yeast DHFR PCA detects direct and near direct interactions without disturbing endogenous regulation, giving insight into the role of transcriptional regulation in the evolution of PPIs (Barshir et al., 2018; Gagnon-Arsenault et al., 2013; Rochette et al., 2014; Tarassov et al., 2008). PCA is one of the standard binary methods used to measure direct and near-direct PPIs in yeast and mammalian cells (Titeca et al., 2019). PCA's performance compares to other standard methods when proper controls and analyses are performed. It has been used successfully by our group and others in various contexts since its first application (Celaj et al., 2017; Chrétien et al., 2018; Lev et al., 2014; Schlecht et al., 2017).

In general, the PCA signal in our study strongly correlates with results from previous PCA experiments (Stynen et al., 2018; Tarassov et al., 2008) and other publicly available data (Figure S1). Roughly 75% of the HMs and HETs detected in our PCA experiments were previously reported (Figure S2), suggesting that most of the HMs and HETs that can be detected with the available tools and in standard conditions have been discovered. While 76 HMs and 47 HETs reported in other studies were not detected in our PCA, our experiments detected 44 HMs and 19 HETs not previously reported.

List of Supplementary Figures

S1 Association between mRNA abundance and the probability of HM detection by PCA in this study. . . . 58
S2 mRNA and protein abundance of singletons and duplicates. . . . 59
S3 Comparison of PCA data generated in this study with published data. . . . 60
S4 Intersections of detected HMs. . . . 61
S5 Interaction motifs and percentage of pairwise amino acid sequence identity between paralogs. . . . 62
S6 Conservation of binding interfaces of human paralogs in HM&HET complexes with solved structures. . . . 63
S7 Plate organization for DHFR PCA experiments. . . . 64
S8 Density of colony size converted to z-score. . . . 65
S9 Comparison of Pfam domain composition similarity between pairs of paralogs. . . . 66
S10 Comparison of functional similarity between HM and HM&HET pairs 67
S11 Comparison of functional similarity between WGDs, considering homeologs and true ohnologs separately. . . . 68
S12 Functional similarity between paralogs as a function of their pairwise amino acid sequence identity. . . . 69
S13 Percentage of interaction motifs for SSDs, WGDs and the two types of WGDs. . . . 70
S14 Similar evolutionary trajectories are observed for six different PDB structures. . . . 71
S15 Effect of changes in parameters on the observed evolution trajectories. . . . 72
S16 Density of colony size converted to z-score 73
S17 Distribution of effect sizes of mutations on the binding energy ($\Delta\Delta$G) of HMs and HETs as estimated using FoldX. . . . 74
S18 Fixation rates of double mutants during the simulations. . . . 75
S19 Contribution of epistasis to the evolution of HET for six different PDB structures. . . . 75
S20 The loss of HETs may result from regulatory divergence (single cell RNAseq data; Gasch et al., 2017). . . . 76
S21 Expression of WGDs and consequences on interaction motifs. . . . 77
S22 Interaction motifs and similarity of functions for SSDs and WGDs. . . . 78
S23 Similarity of regulation between paralogs as a function of their pairwise amino acid sequence identity. . . . 78

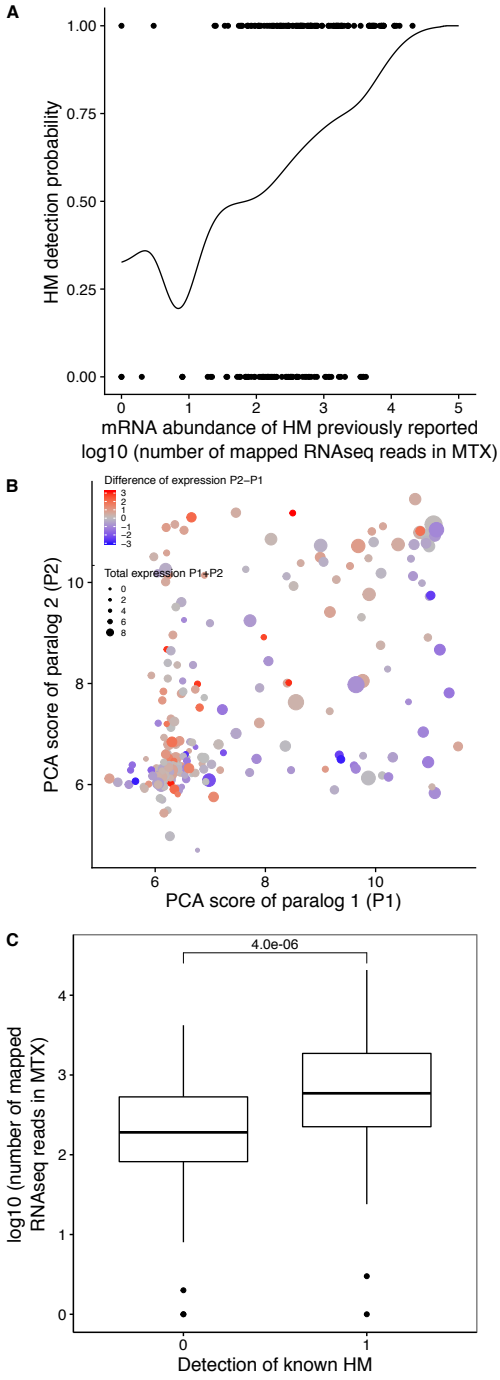

Figure S1: Association between mRNA abundance and the probability of HM detection by PCA in this study.
(A) The probability that PCA detects a HM is correlated with expression level, as estimated by RNAseq. The plot shows the detection probability of HMs as a function of mRNA abundance for previously reported HMs. Kernel regression of the HM detection (one for detected, 0 for not detected) on the number of mapped reads per gene (log10). **(B)** Difference in HM formation between paralogs results in part from their differential mRNA abundance. The PCA score of paralog 1 (P1) is compared to the PCA score of paralog 2 (P2). PCA scores are median colony sizes from the PCA experiments performed in this study. The total mRNA abundance of paralogs is shown by the size of the points and the difference of expression levels is represented by a color gradient (red for overexpression of P2 compared to P1 and blue overexpression of P1 compared to P2). Red points tend to be above the diagonal, blue points, below the diagonal. **(C)** Comparison of expression levels of previously reported HMs for HMs undetected and detected in the PCA experiments performed in this study. P-value from a Wilcoxon test is shown. Comparison of PCA data generated in this study with previously published data.

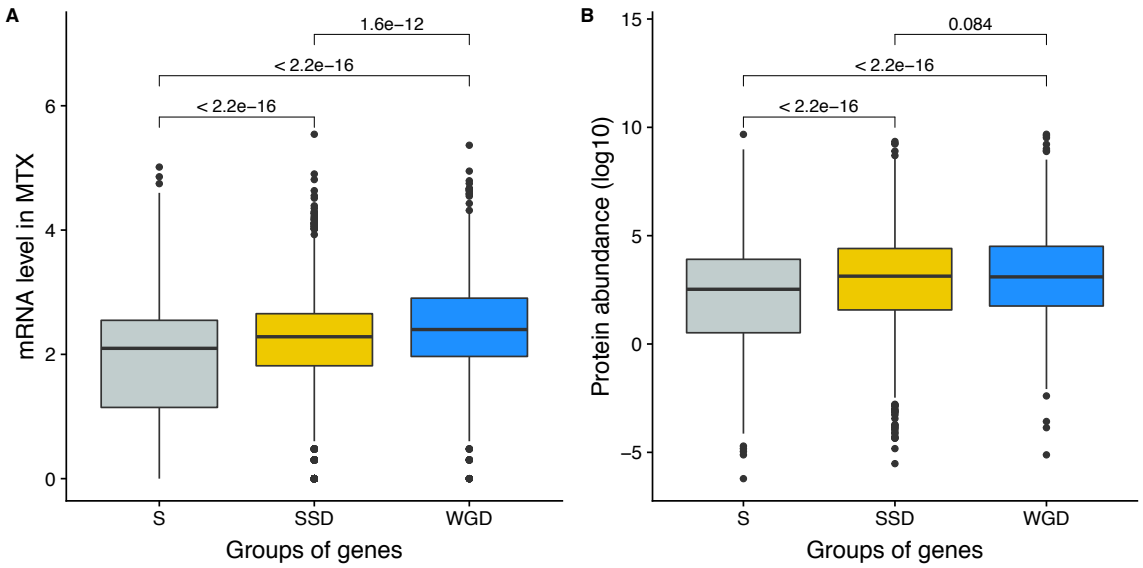

Figure S2: mRNA and protein abundance of singletons and duplicates.
(A) Comparison of mRNA abundance of genes as a function of whether they rare duplicated and of their type of duplication. **(B)** Comparison of the protein abundance as a function of whether they rare duplicated and their type of duplication. (S: singleton, SSD: Small-Scale Duplicates, WGD: Whole-Genome Duplicates). Numbers indicate p-values from Wilcoxon tests.

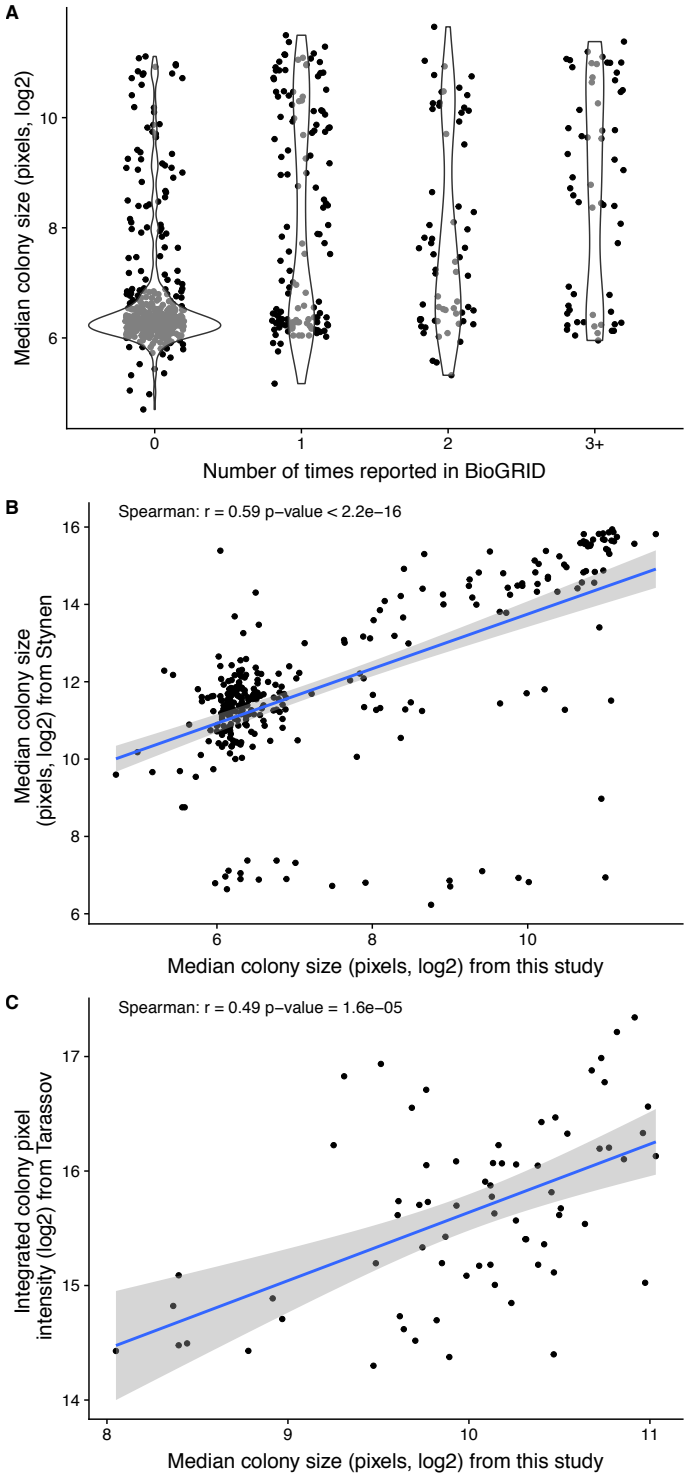

Figure S3: Comparison of PCA data generated in this study with published data.
(A) Colony size (estimated as the integrated pixel intensity) in the PCA experiment as a function of the number of times the corresponding interaction is reported in BioGRID version BIOGRID-3.5.166 (Chatr-Aryamontri et al., 2012, 2017). **(B)** Correlation between colony size of the study of Stynen et al. (2018) on homomers and of the PCA experiment performed in this study. **(C)** Correlation between colony size of Tarassov et al. (2008) and of the PCA experiment performed in this study.

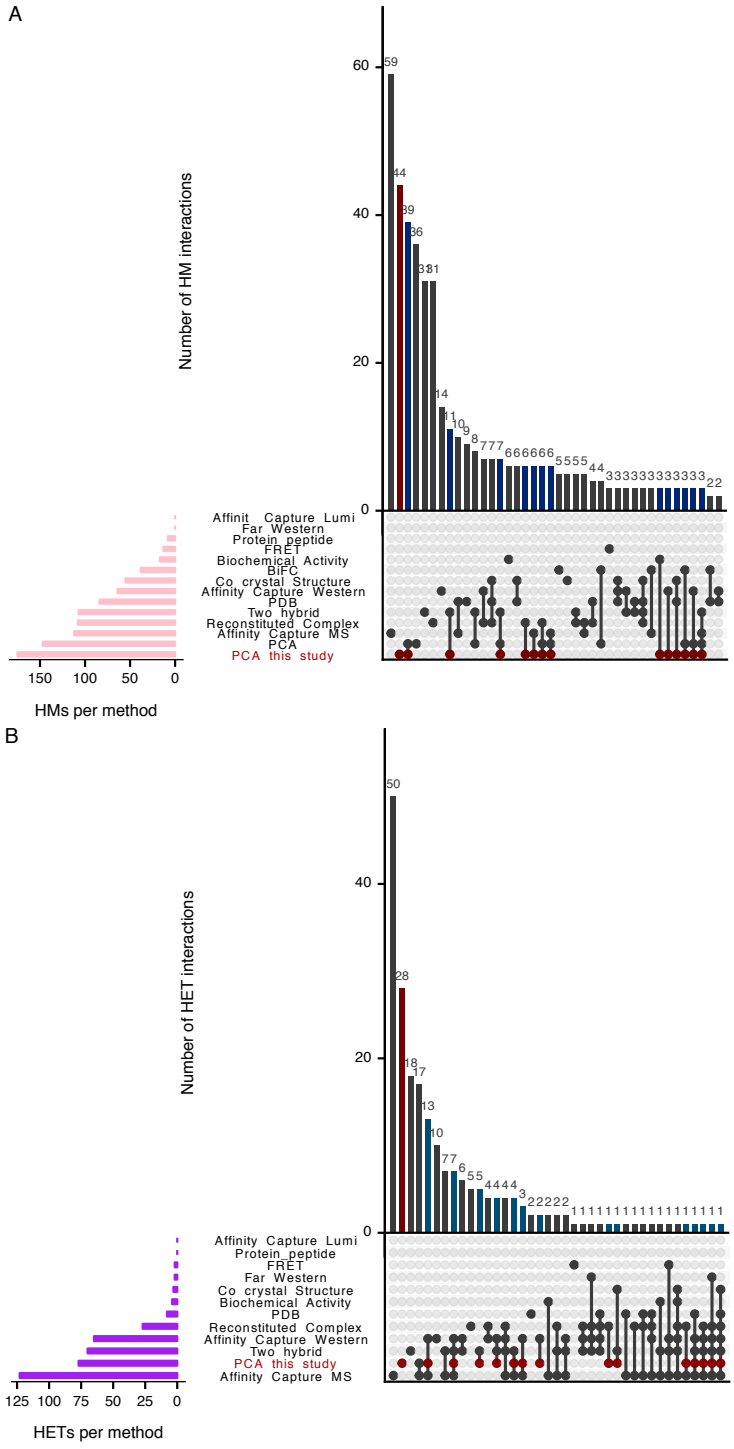

Figure S4: Intersections of detected HMs.
(A) and HETs **(B)** from this study and previously reported HMs and HETs. We considered HMs and HETs reported in crystal structures from the Protein Data Bank on September 21st, 2017 (Berman et al., 2000) and by PCA based on fluorescent proteins (BiFC) (Kim et al., 2019). We also include HMs and HETs reported in BioGRID (BIOGRID-3.5.166; Chatr-Aryamontri et al., 2013; Chatr-Aryamontri et al., 2017) with these methods: Affinity Capture-MS, Affinity Capture-Western, Reconstituted Complex, Two-hybrid, Biochemical Activity, Co-crystal Structure, Far Western, FRET, Protein-peptide, Affinity Capture-Luminescence and PCA. We added data from Stynen et al. (2018) to the BioGRID PCA data. Results of the PCA experiments from this study are highlighted in red. Turquoise-blue bars show HMs and HETs detected in this study and previously observed. The intersections were computed and plotted using the R package UpSetR (Lex et al., 2014).

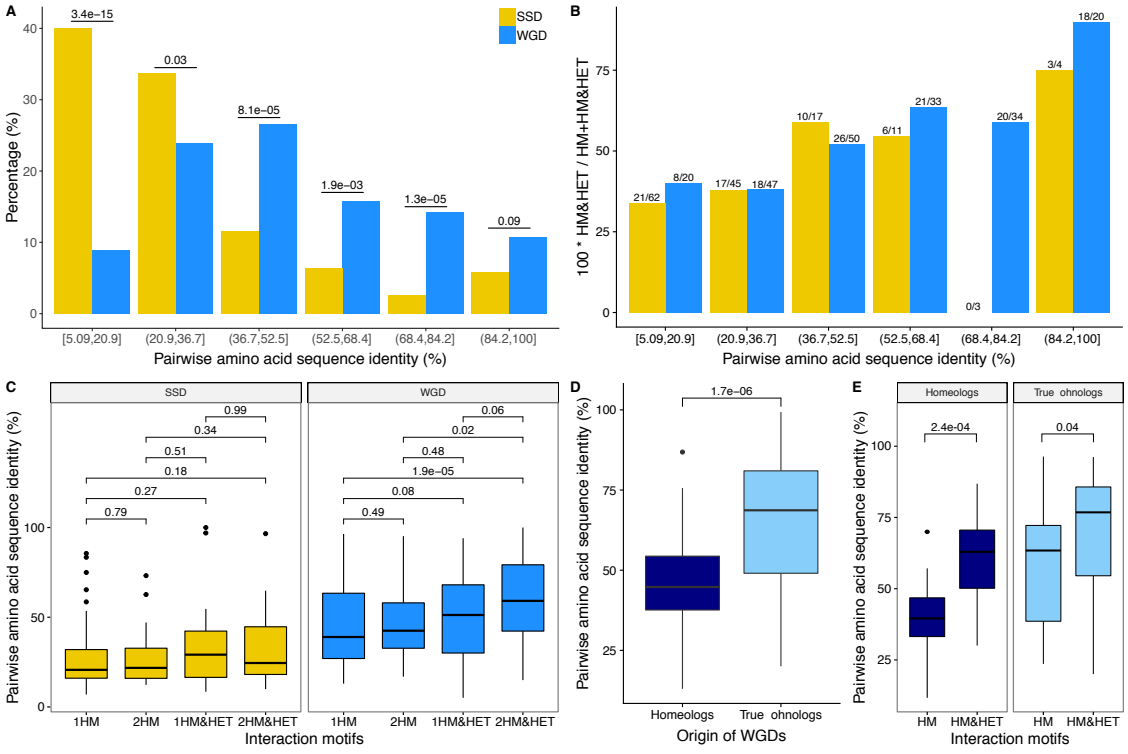

Figure S5: Interaction motifs and percentage of pairwise amino acid sequence identity between paralogs. **(A)** Pairs of paralogs were clustered in six pairwise amino acid sequence identity groups and the distribution (in percentage) of these groups were compared between SSD and WGD. P-values are from Fisher's exact tests. **(B)** The percentage of paralog pairs forming HM&HET among the total number of paralog pairs forming at least one HM (HM and HM&HET) is shown as a function of the percentage of pairwise amino acid sequence identity (SSDs in yellow and WGDs in blue). For each group, the number of HM&HET pairs and the total number are indicated above the bars. **(C)** Percentage of pairwise amino acid sequence identity between paralogs for each motif. 1HM: shows one homomer only, 2HM: shows both homomers, 1HM&HET: shows one homomer and the heteromer, and 2HM&HET: shows both homomers and the heteromer. P-values are from Wilcoxon tests. **(D)** The percentage of pairwise amino acid sequence identity among homeologs (dark blue) and true onhologs (light blue). P-value is from a Wilcoxon test. **(E)** Percentage of pairwise amino acid sequence identity between paralogs for HM and HM&HET motifs for homeologs and true ohnologs. P-values are from Wilcoxon tests.

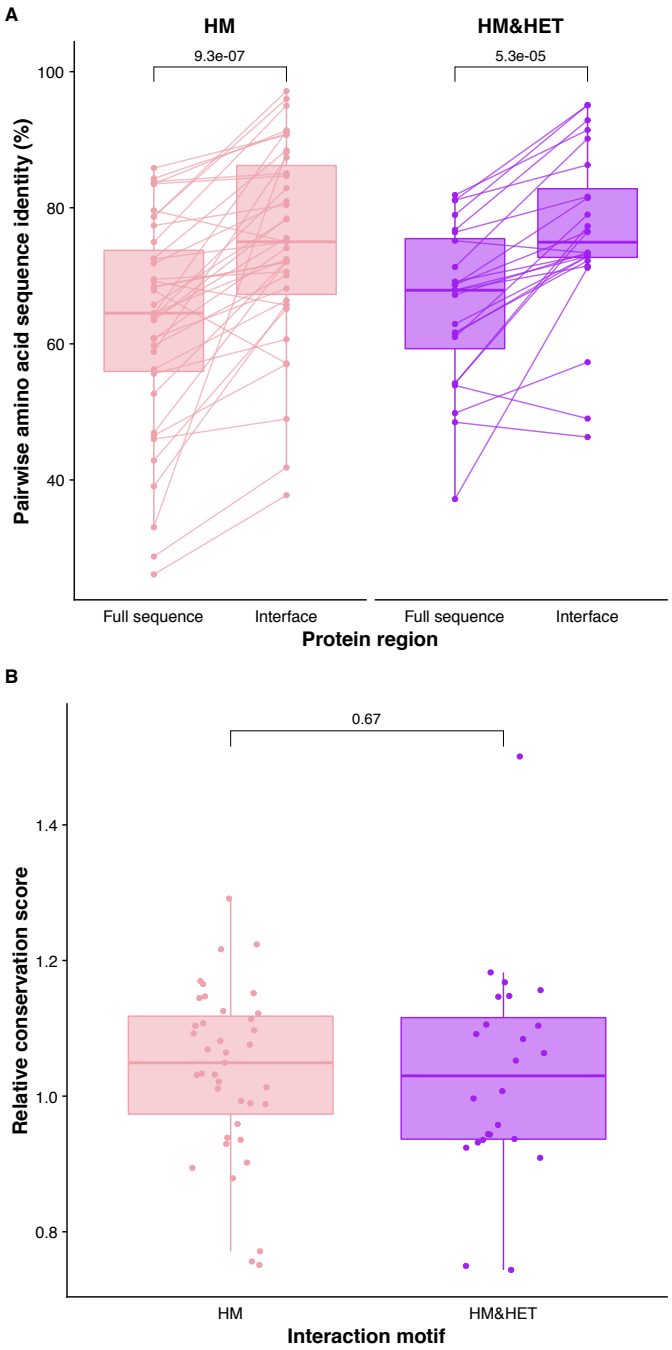

Figure S6: Conservation of binding interfaces of human paralogs in HM&HET complexes with solved structures. (A) Pairwise amino acid sequence identity for the full sequences of paralogs and their interfaces are shown for the two motifs). P-values from paired Wilcoxon tests are shown. **(B)** Relative conservation scores are shown for the two motifs of paralogs. Relative conservation scores are calculated based on the protein regions solved by crystallography as the percentage of sequence identity at the binding interface divided by the percentage of sequence identity outside the interface. Paralog pairs were classified as HM or HM&HET. Homologous interfaces were identified in alignments of the paralogous sequences. The list of PDB IDs used for these analyses includes 40 interfaces from homomeric structures for the HM group and 25 interfaces for the HM&HET group (24 homomers and 1 heterodimer of paralogs). P-value is from a Wilcoxon test.

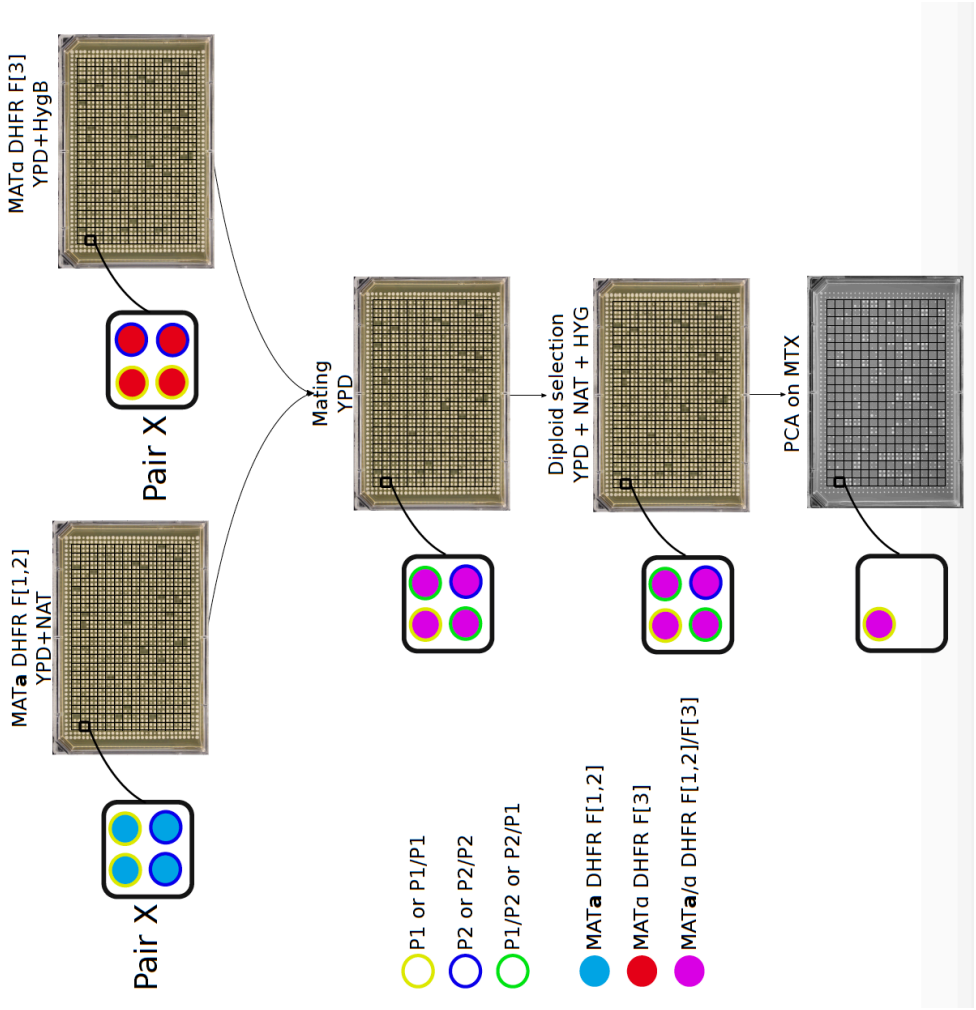

Figure S7: Plate organization for DHFR PCA experiments.
On the haploid arrays (MAT**a** and MATα), each plate has two rows and two columns of control strains at the border (blue lines). Paralogs of a pair are positioned in blocks of four strains. A given pair (example here of pair X) occupies the same position in the MAT**a** and MATα plates. Inside a square, paralogs are positioned horizontally in MATa DHFR F[1,2] plates (P1 are at the top and P2 at the bottom of the square) while they are vertically positioned in MATα DHFR F[3] plates (P1 are at the left and P2 at the right of the square). The two haploid plates were printed on top of each other on a mating plate, generating the following crosses: P1-DHFR F[1,2]/P1 DHFR F[3] at top left, P1-DHFR F[1,2]/P2 DHFR F[3] at top right, P2-DHFR F[1,2]/P1 DHFR F[3] at bottom left and P2-DHFR F[1,2]/P2 DHFR F[3] at bottom right. Two diploid selections and two replications on MTX medium were performed.

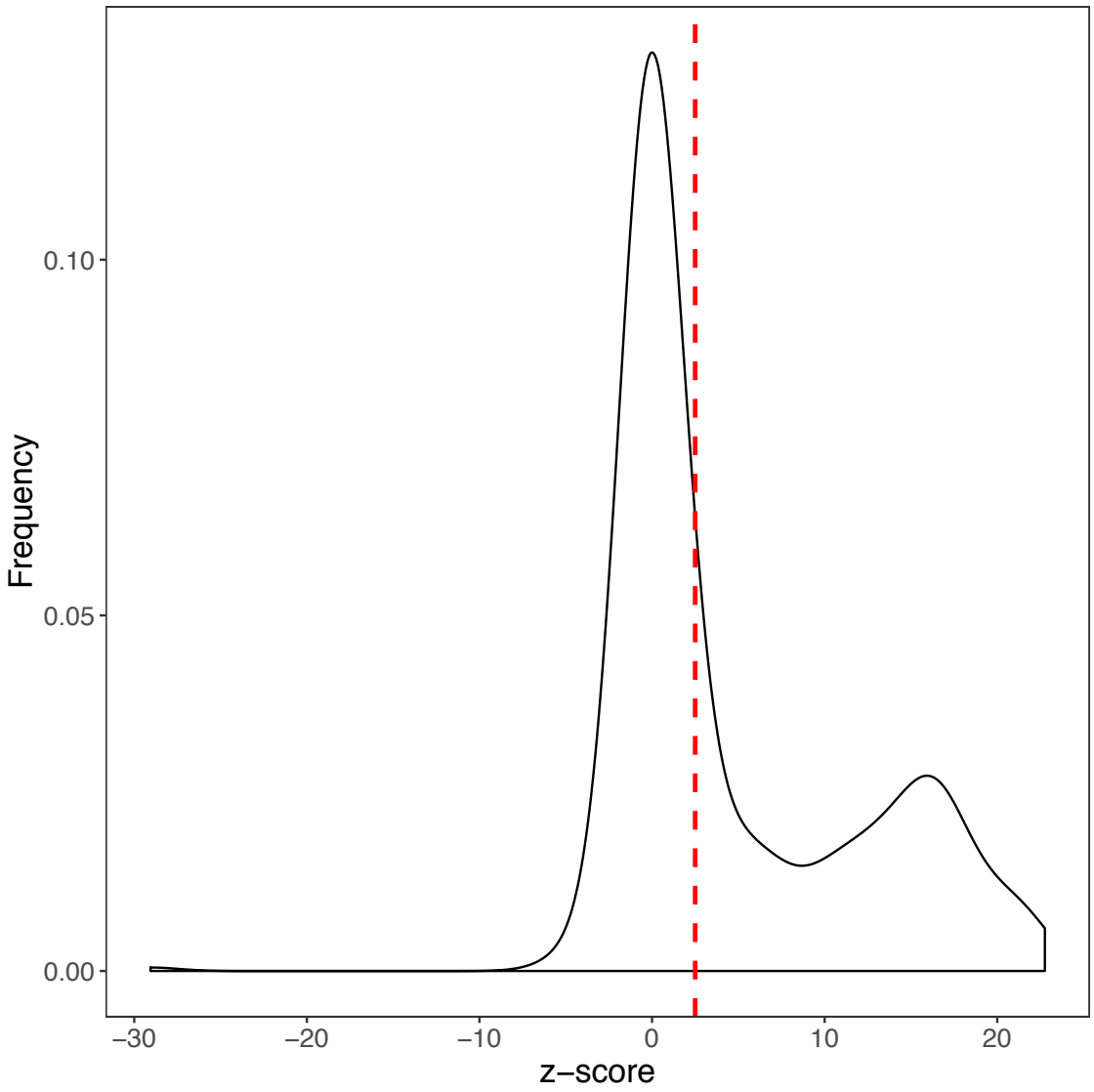

Figure S8: Density of colony size converted to z-score.
Colony sizes from the PCA experiment of this study were converted to z-score using the mean (μ_b) and standard deviation (sd_b) of the background distribution ($Z_s = (I_s - \mu_b)/sd_b$)). The density of z-scores is shown in black. A protein-protein interaction was considered as detected if the corresponding z-score was larger than 2.5 (red dashed line).

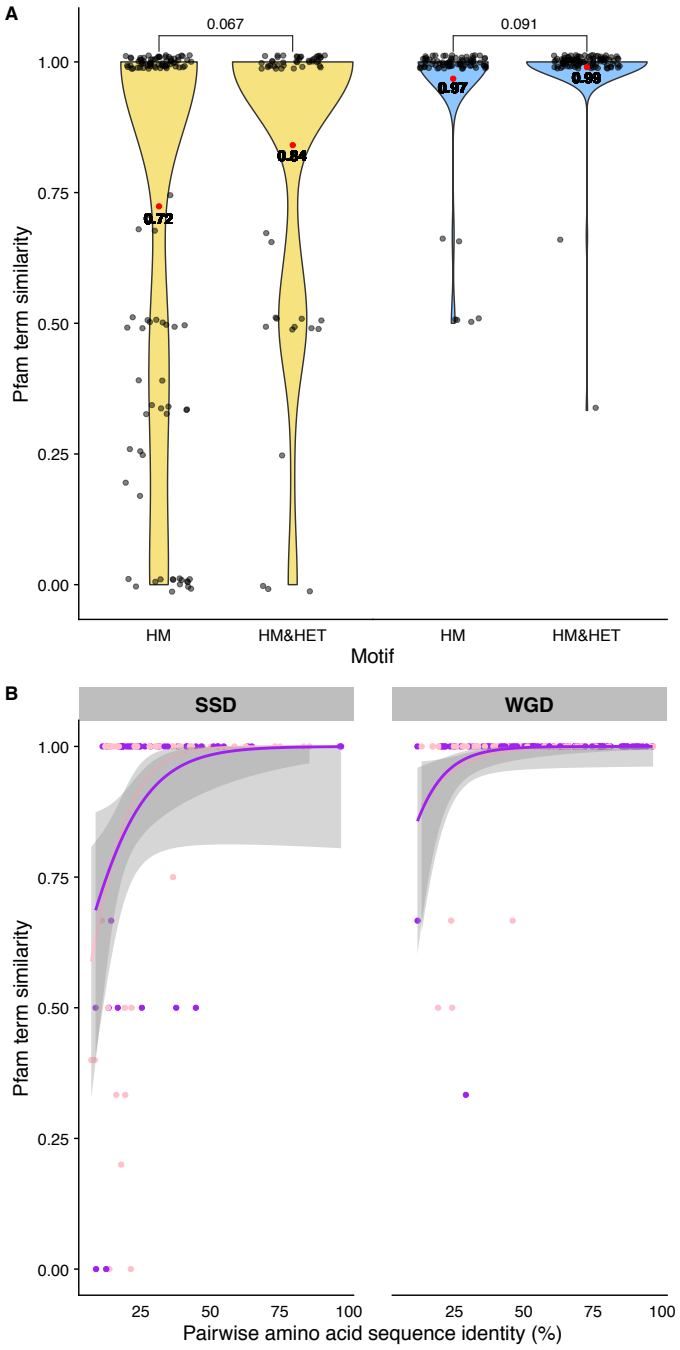

Figure S9: Comparison of Pfam domain composition similarity between pairs of paralogs.
(A) Pfam domain composition similarity (Jaccard's index) between SSDs (yellow) and WGDs (blue) for each interaction motif (HM or HM&HET). **(B)** Pfam domain composition similarity as a function of pairwise amino acid sequence identity for HM motifs (pink) and HM&HET motifs (purple). Regression lines were smoothed using the GLM function with the quasibinomial family.

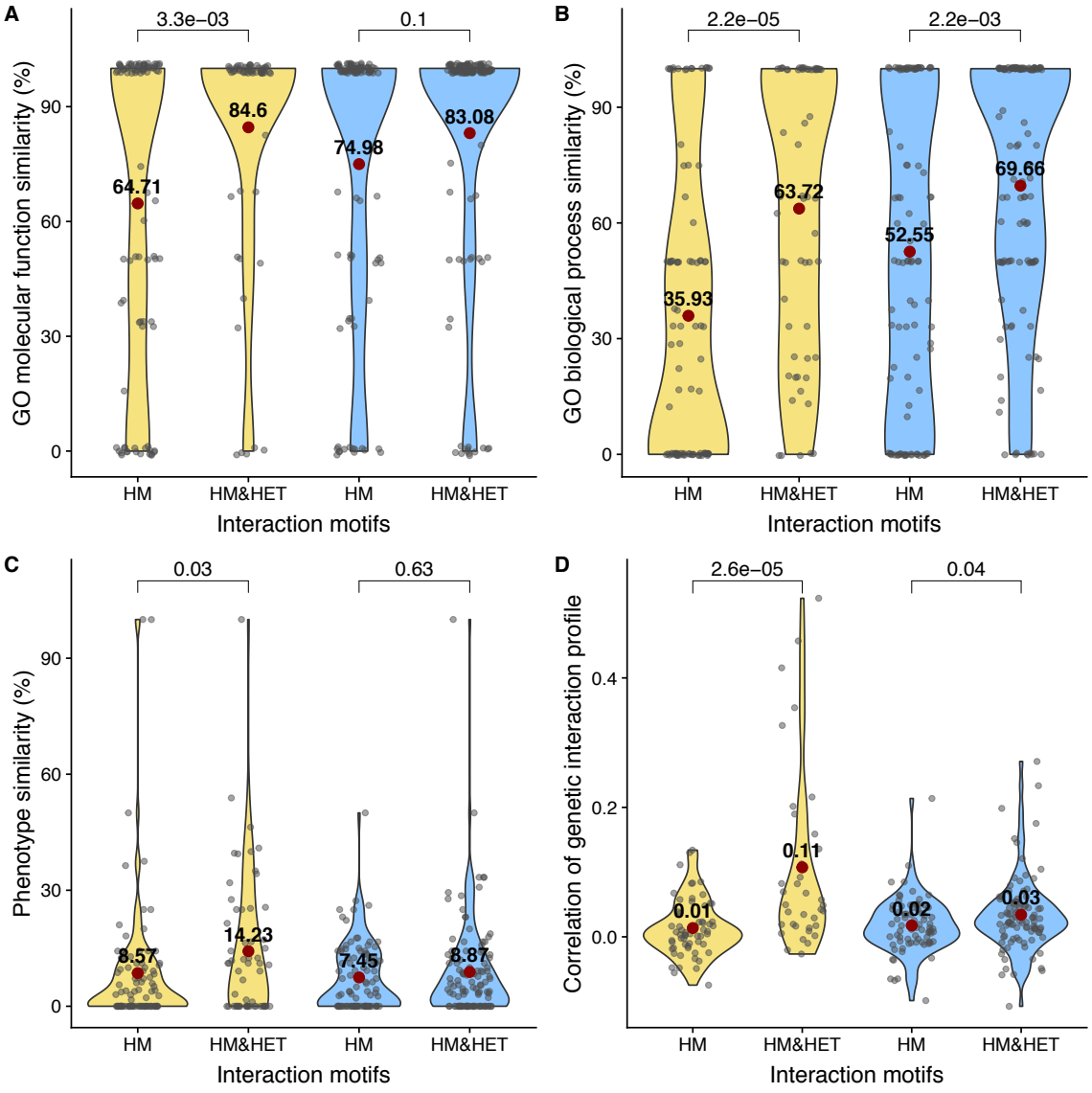

Figure S10: Comparison of functional similarity between HM and HM&HET pairs.
The similarity of function (100% * Jaccard's index) between SSDs (yellow) and WGDs (blue) was estimated using GO terms for *(A)* molecular functions and for *(B)* biological processes. The similarity of function was also estimated using *(C)* growth phenotypes and *(D)* the correlation of genetic interaction profiles. P-values are from Wilcoxon tests.

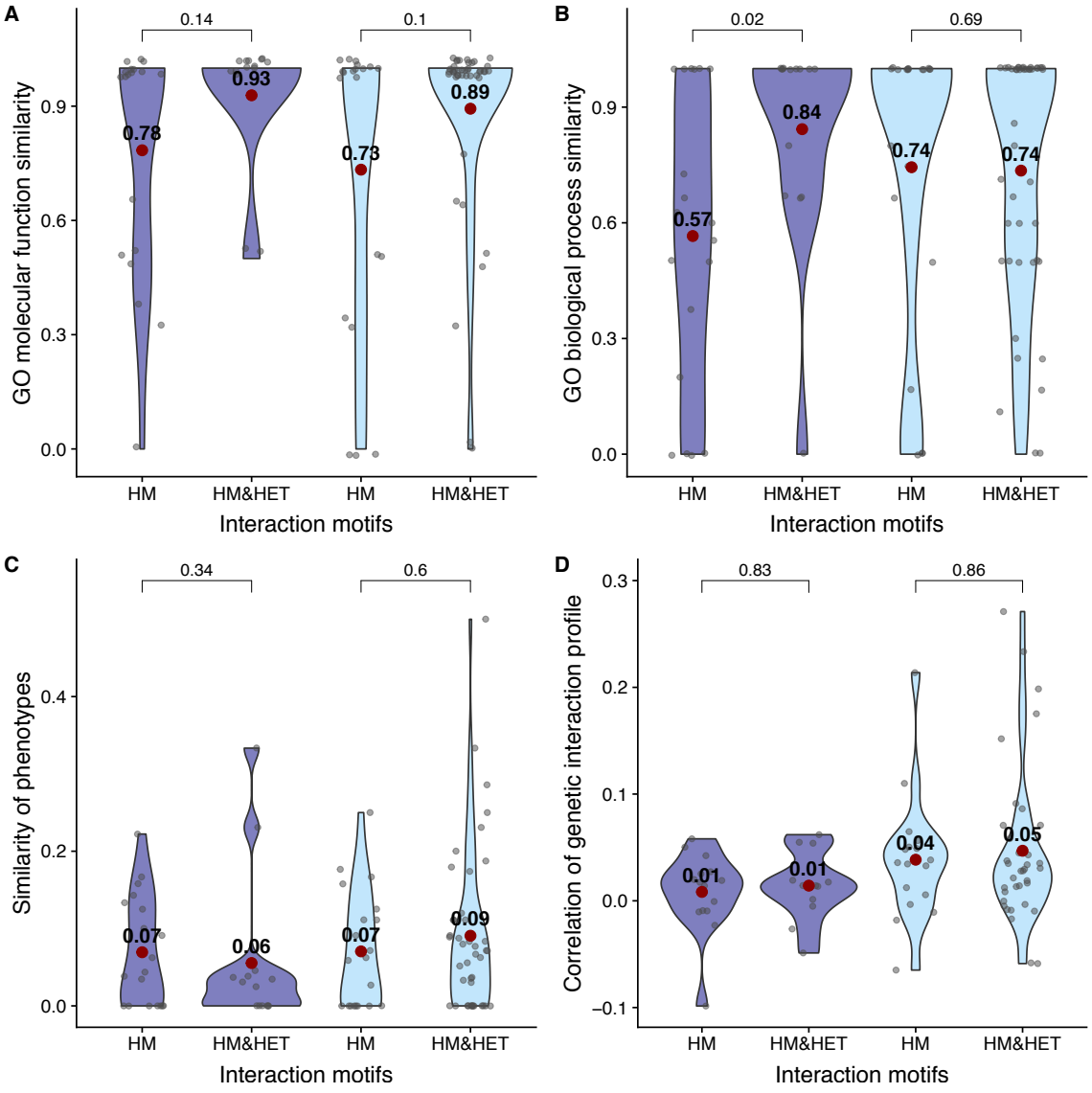

Figure S11: Comparison of functional similarity between WGDs, considering homeologs and true ohnologs separately.
The similarity of function (100% * Jaccard's index) between homeologs (dark blue) and true ohnologs (light blue) was estimated using GO terms for **(A)** molecular functions and for **(B)** biological processes. The similarity of functions was also estimated using **(C)** growth phenotypes and **(D)** the correlation of genetic interaction profiles. P-values are from Wilcoxon tests.

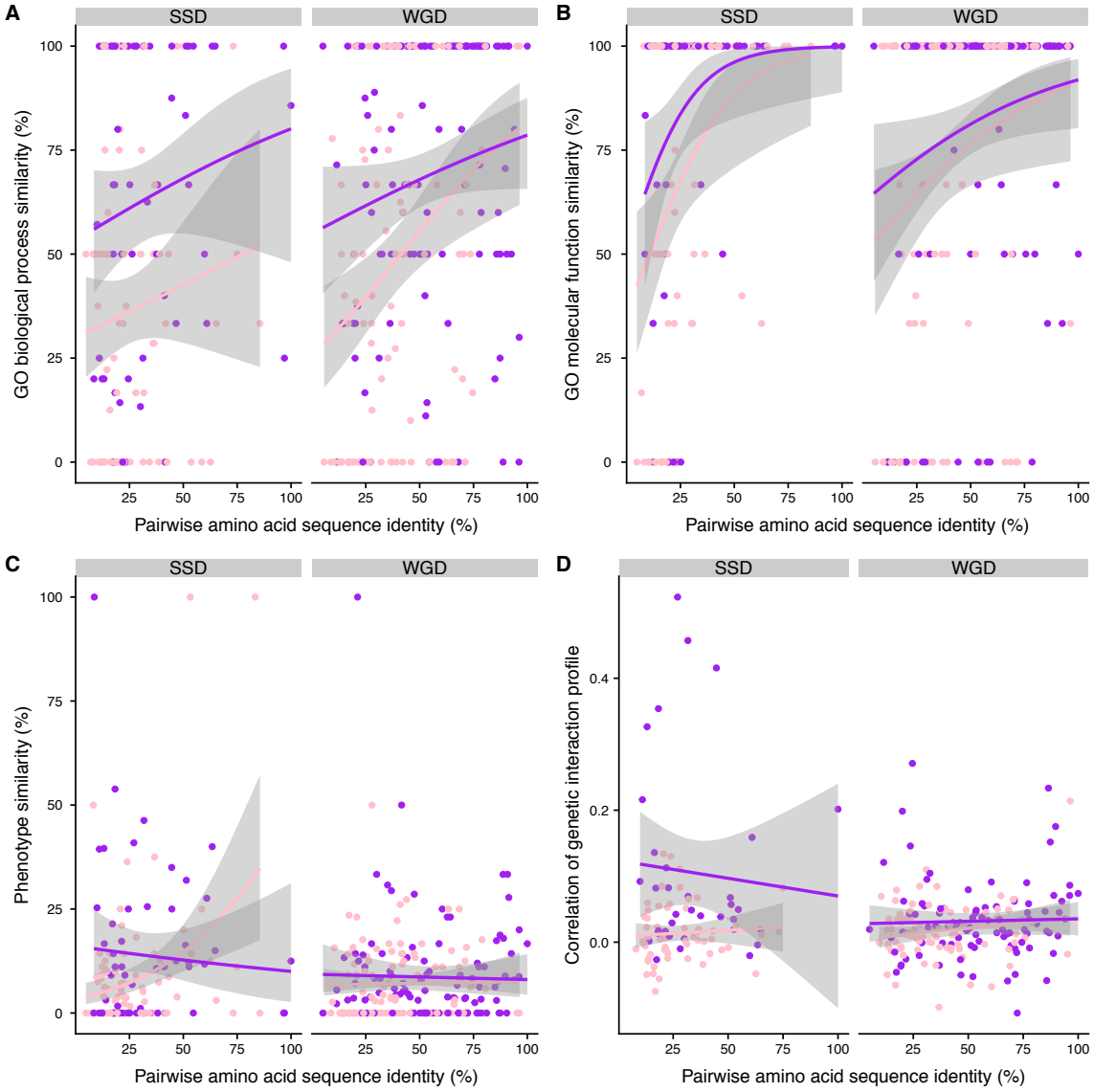

Figure S12: Functional similarity between paralogs as a function of their pairwise amino acid sequence identity.
The similarity of function (100% * Jaccard's index) between paralogs for HM (pink) and HM&HET (purple) as a function of pairwise amino acid sequence identity for SSDs and WGDs. Similarity of function was estimated using **(A)** molecular functions and **(B)** biological processes GO terms, **(C)** growth phenotypes and **(D)** the correlation of genetic interaction profiles. The regression lines were smoothed using the R geom_smooth function.

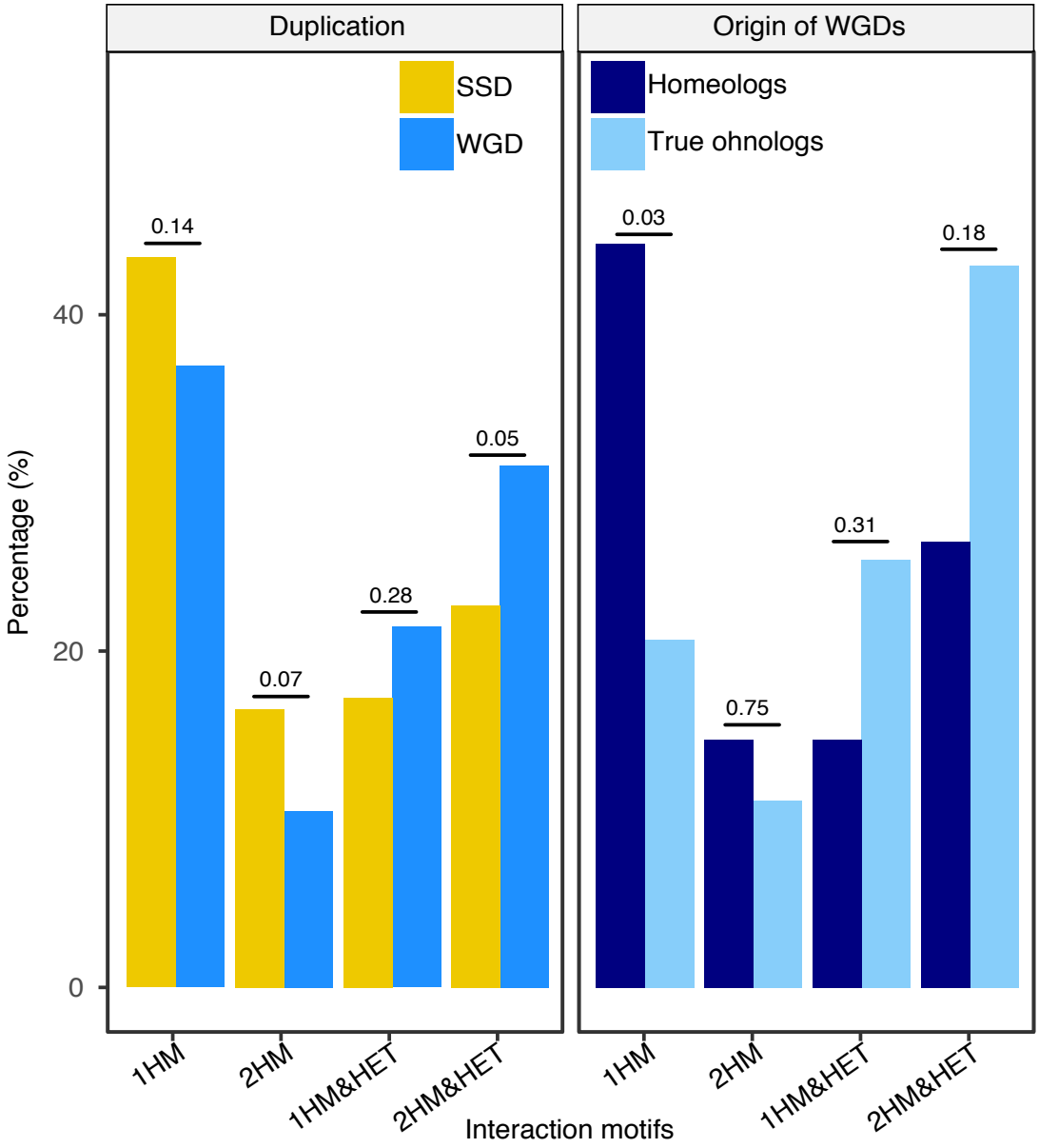

Figure S13: Percentage of interaction motifs for SSDs, WGDs and the two types of WGDs. The data is the same as shown in Figure 2 but all four possible HM and HM&HET motifs are shown. 1HM: shows one homomer only, 2HM: shows both homomers, 1HM&HET: shows one homomer and the heteromer and 2HM&HET: shows both homomers and the heteromer. The percentage of motifs of interaction for SSDs (yellow) and WGDs (blue) (left panel) and for homeologs (dark blue) and true ohnologs (light blue) (right panel). P-values are from Fisher's exact tests.

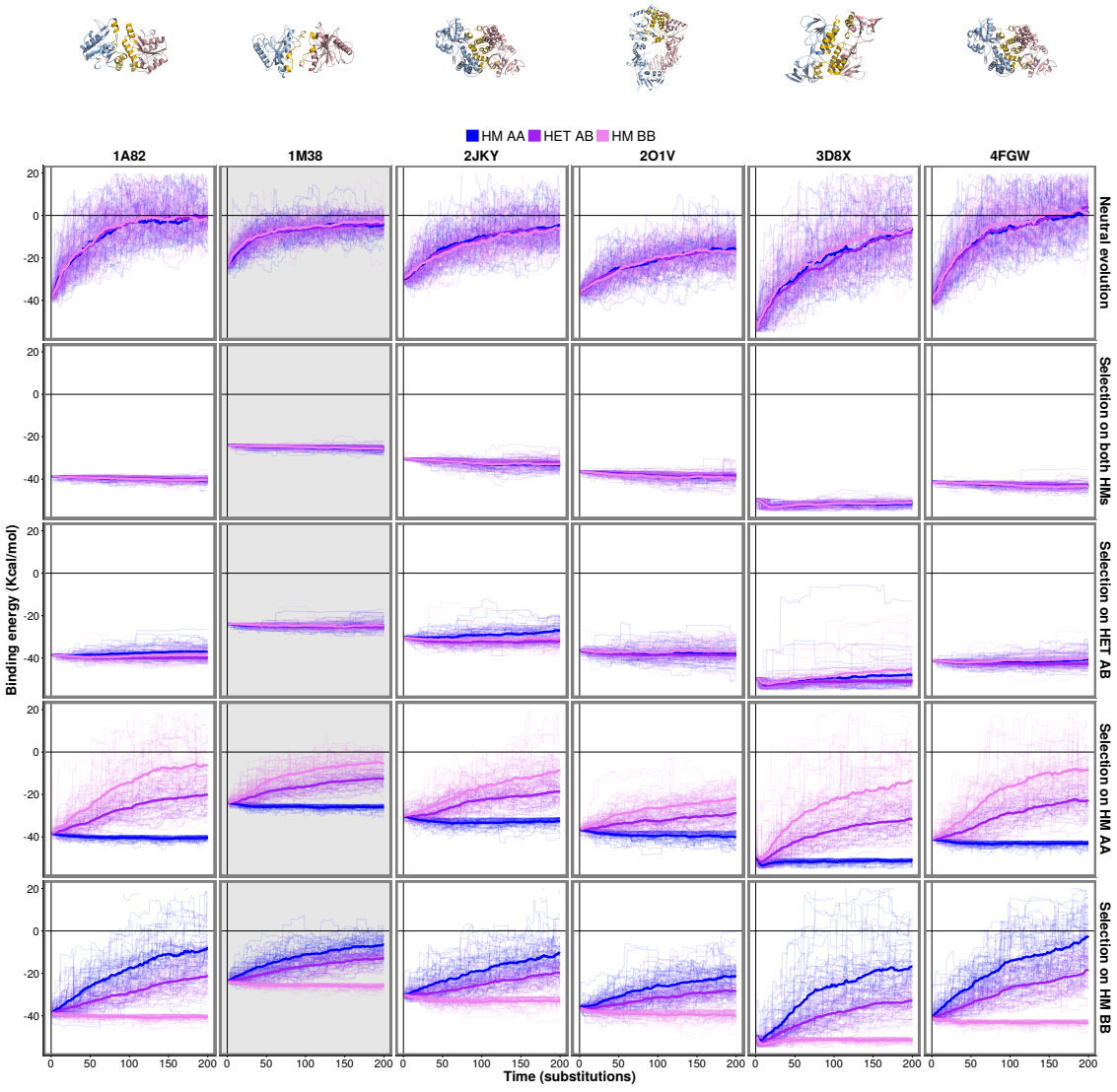

Figure S14: Similar evolutionary trajectories are observed for six different PDB structures.
The binding energy of six HMs and HETs is followed through time under the same scenarios as shown in Figure 4. Panels shown in Figure 4 are highlighted with a gray background here.

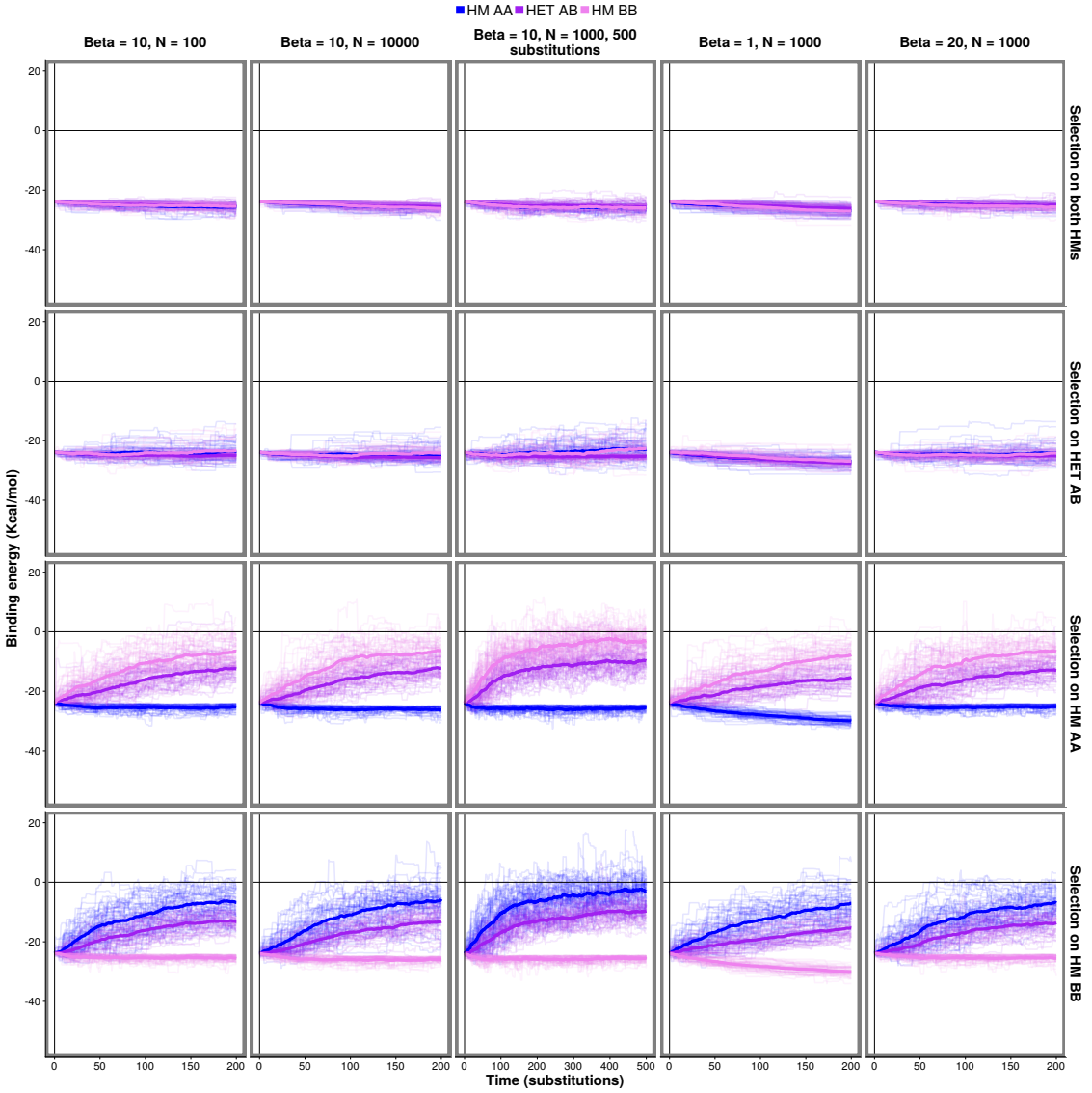

Figure S15: Effect of changes in parameters on the observed evolution trajectories.
Simulations were run for different combinations of parameters controlling the efficiency of selection (β and N) and the length of the simulations for PDB structure 1M38.

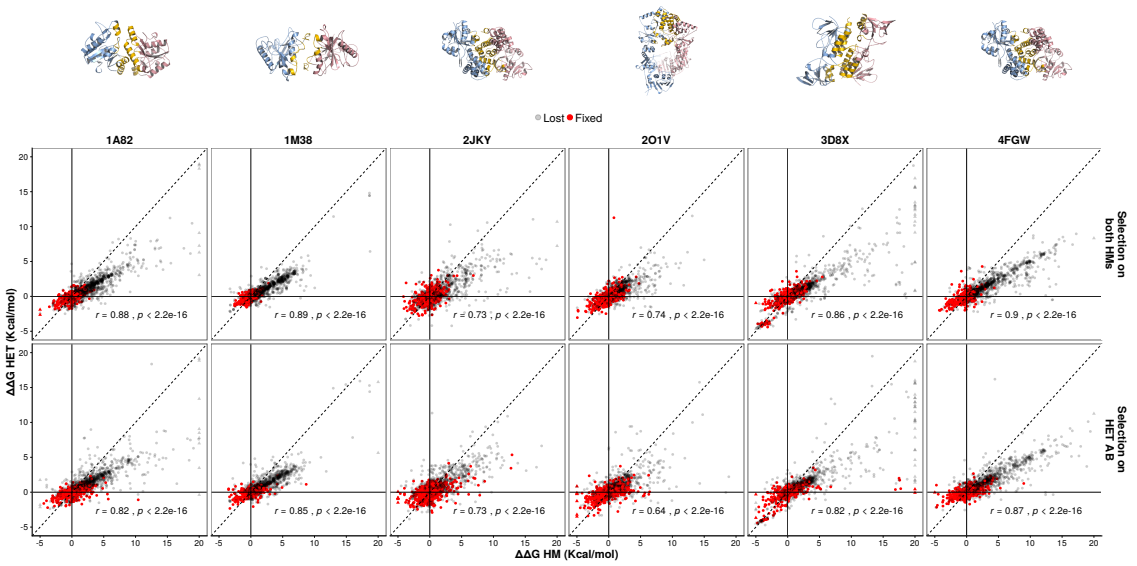

Figure S16: Single mutants have pleiotropic effects for HM and HET.
The observed effects of sampled single mutants on the HET are compared with their effects on HMs. Pearson's correlation coefficients are shown. Parameters used for β and N were 10 and 1000, respectively.

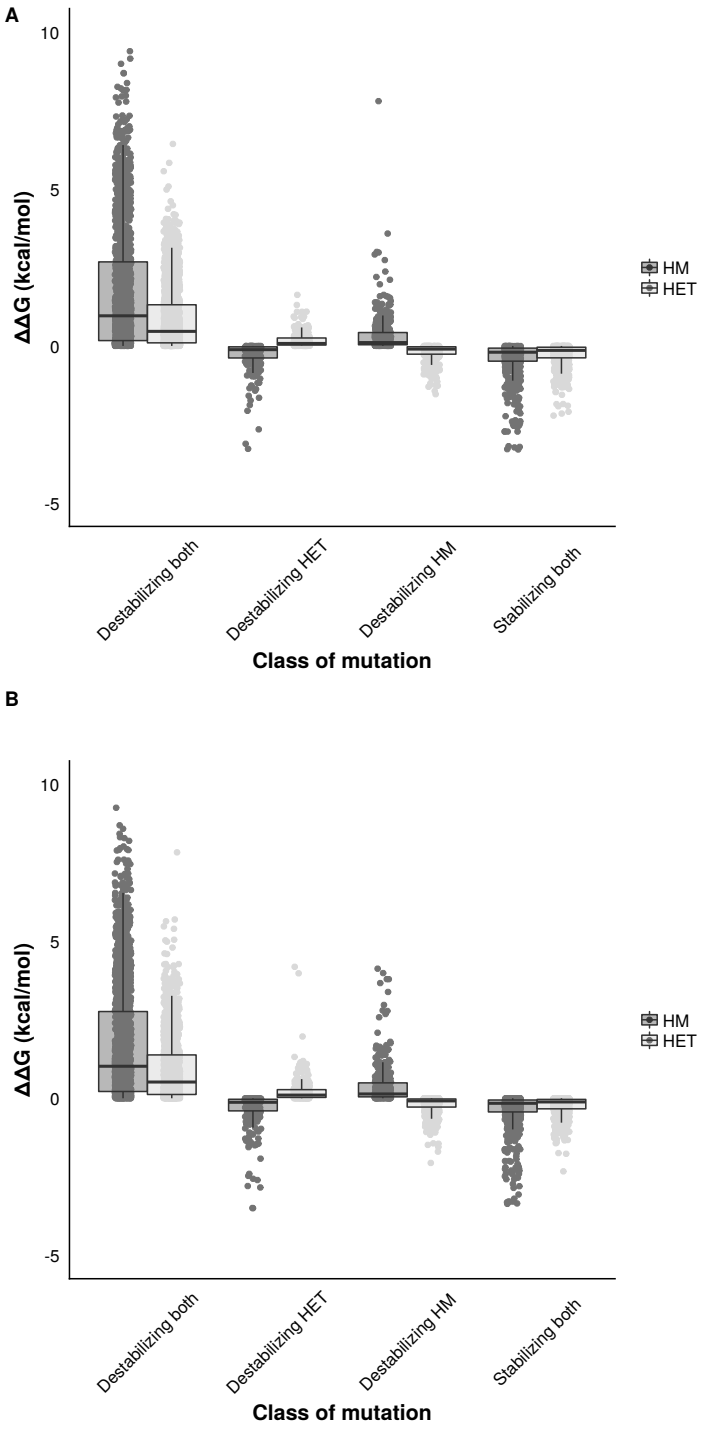

Figure S17: Distribution of effect sizes of mutations on the binding energy ($\Delta\Delta$G) of HMs and HETs as estimated using FoldX.
Effects of single mutants on the binding energy of HMs and HETs. Mutants were classified (x-axis) according to their effects on the binding energy of HMs and HETs, depending on whether they stabilize or destabilize both the HM and the HET or they only destabilized one of them. Mutations that destabilize one of the complexes have smaller effect sizes on binding energy than mutations that destabilize or stabilize both. **(A)** Mutations sampled when negatively selecting for the stability of both HMs. **(B)** Mutations sampled when negatively selecting for the stability of the HET. Parameters used for β and N were 10 and 1000, respectively.

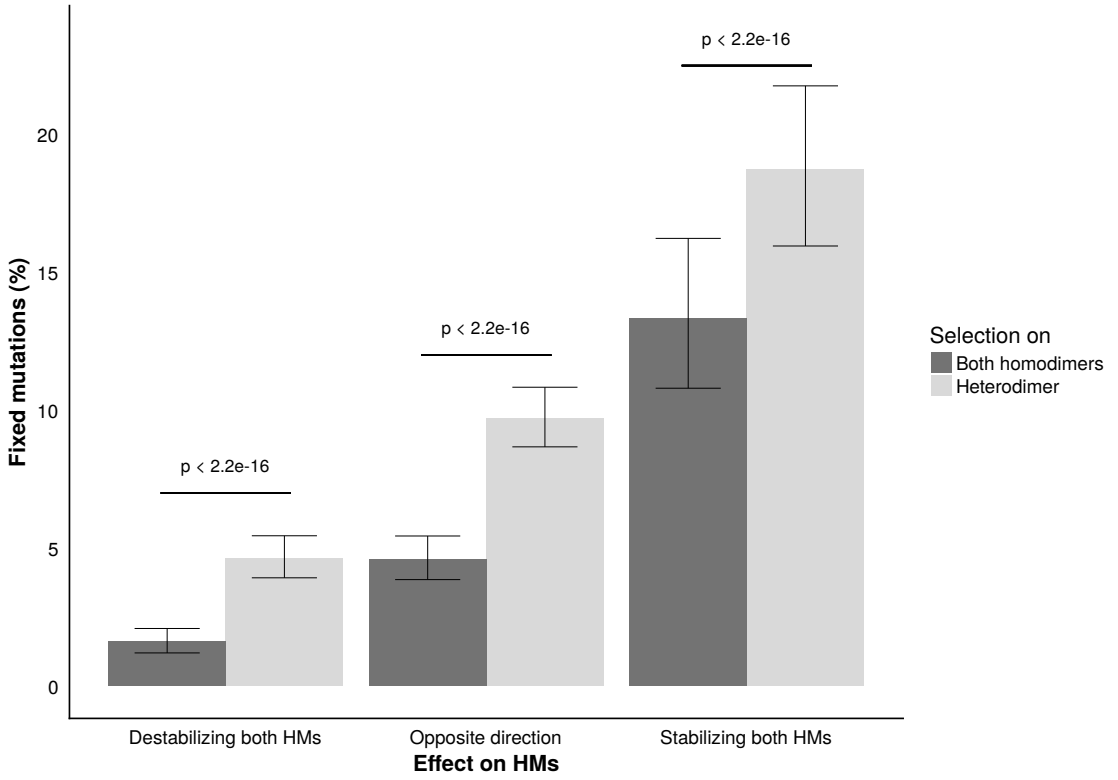

Figure S18: Fixation rates of double mutants during the simulations.
Fixation rates of double mutants classified based on their effect on the two HMs and the complexes (both HMs or HET) under selection. Clopper-Pearson 95% confidence intervals are shown. P-values were calculated with a two proportion z-test. Parameters used for β and N were 10 and 1000, respectively.

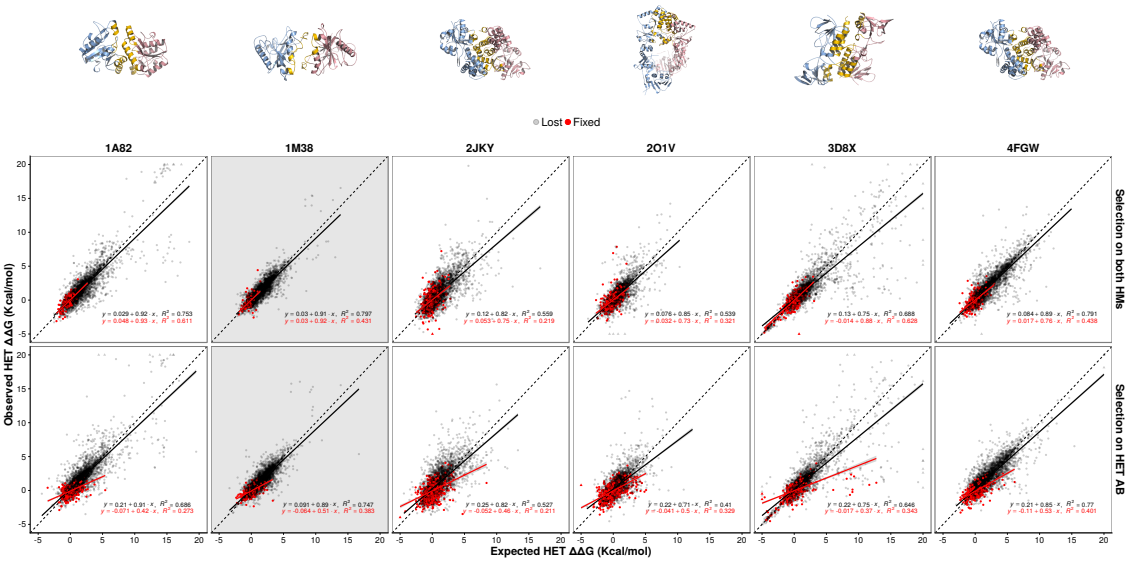

Figure S19: Contribution of epistasis to the evolution of HET for six different PDB structures.
The observed effects of double mutants on the HET are compared with their expected effects based on the effects on the HMs throughout the simulations. Simulations were run under the same scenarios shown in Figure 5. Panels shown in Figure 5 are highlighted with a gray background. Red points are for mutations that were fixed, gray ones those that were eliminated by selection. The regression equations are shown for fixed and lost mutations separately. Parameters used for β and N were 10 and 1000, respectively.

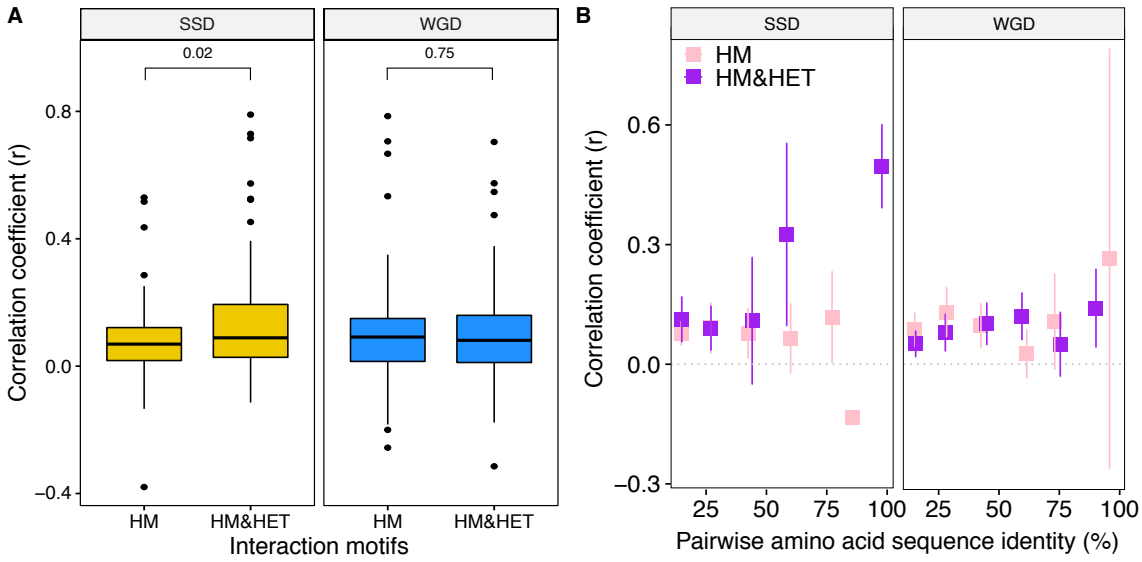

Figure S20: The loss of HETs may result from regulatory divergence (single cell RNAseq data; Gasch et al., 2017).
(A) Correlation (Spearman's r) between the expression profile of paralogs are compared among the different interaction motifs for SSDs (yellow) and WGDs (blue). P-values are from t-tests. **(B)** Correlation of expression profiles between paralogs forming only HM (pink) or HM&HET (purple) as a function of their pairwise amino acid sequence identity.

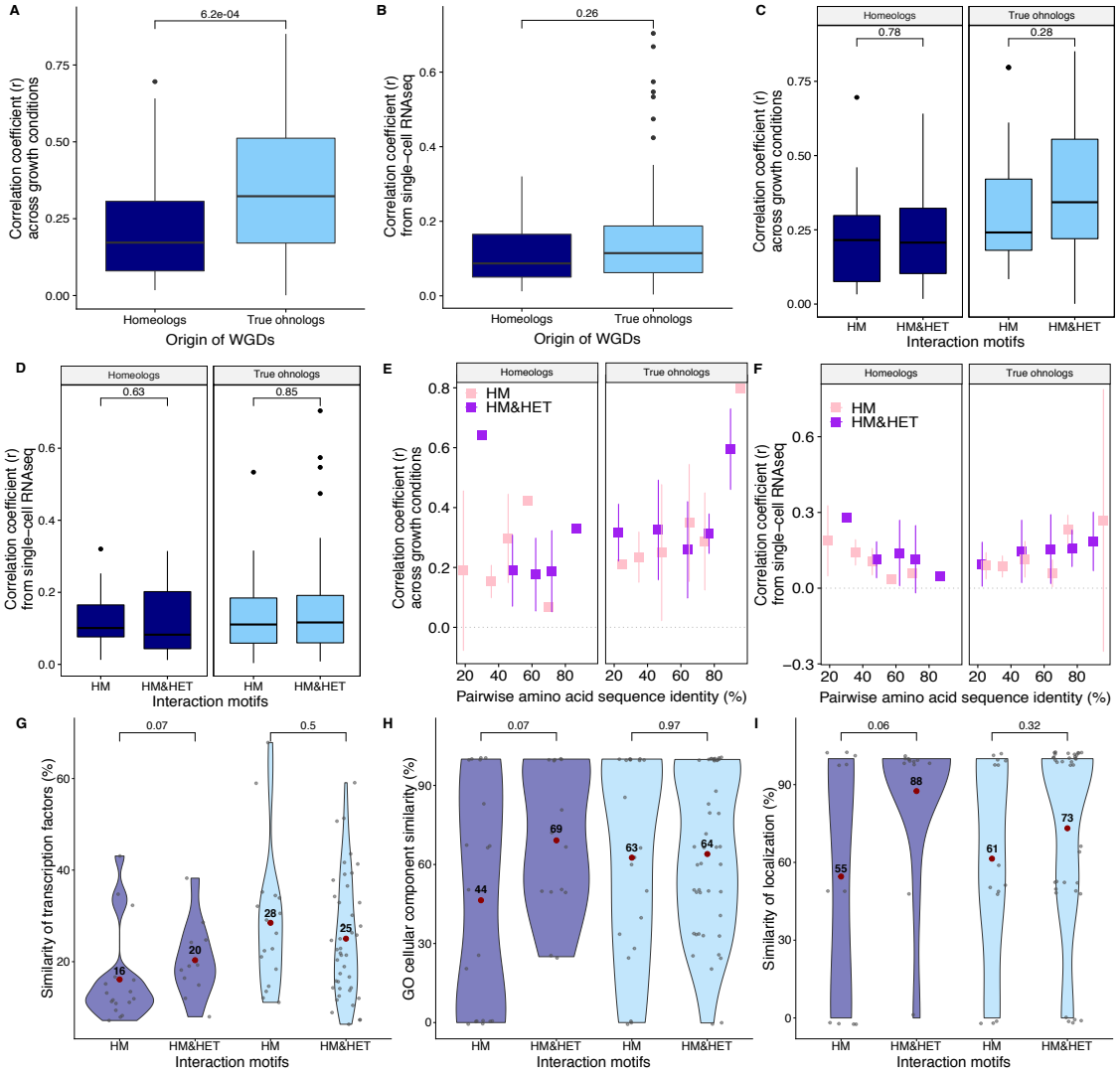

Figure S21: Expression of WGDs and consequences on interaction motifs.
Correlation coefficients (Spearman's r) between the expression profiles of paralogs **(A)** from mRNA relative expression across 1000 growth conditions (Ihmels et al., 2004) and **(B)** from single-cell RNAseq (Gasch et al., 2017) are compared between homeologs and true ohnologs. Correlation coefficients (Spearman's r) **(C)** across growth conditions and **(D)** from single-cell RNAseq data (Gasch et al., 2017) are compared among the different interaction motifs for homeologs and true ohnologs. Correlation coefficients **(E)** across growth conditions and **(F)** from single-cell RNAseq as a function of the percentage of pairwise amino acid sequence identity between paralogs forming only HM or HM&HET. **(G)** Similarity of transcription factor binding sites (100% * Jaccard's index). **(H)** Similarity of GO cellular components. **(I)** Similarity of localization. P-values are from Wilcoxon tests.

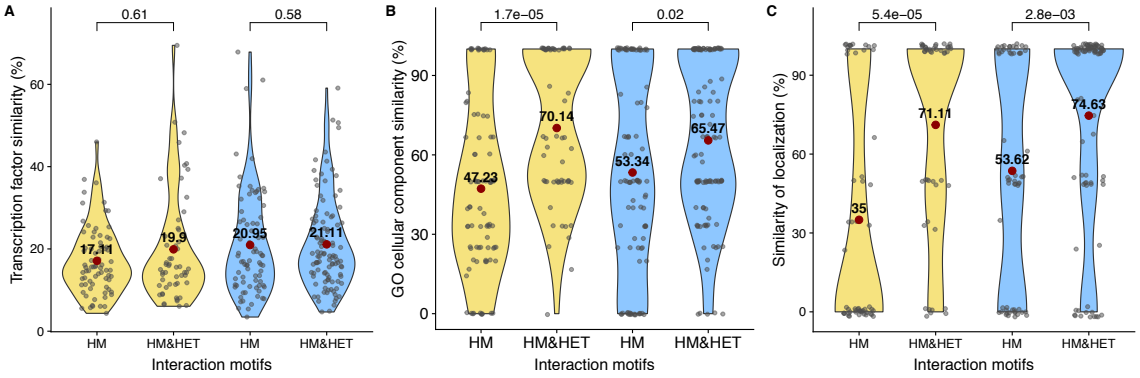

Figure S22: Interaction motifs and similarity of functions for SSDs and WGDs.
The similarity of regulation (100% * Jaccard's index) for **(A)** transcription factor binding sites, **(B)** GO cellular components and **(C)** localization. P-values are from Wilcoxon tests.

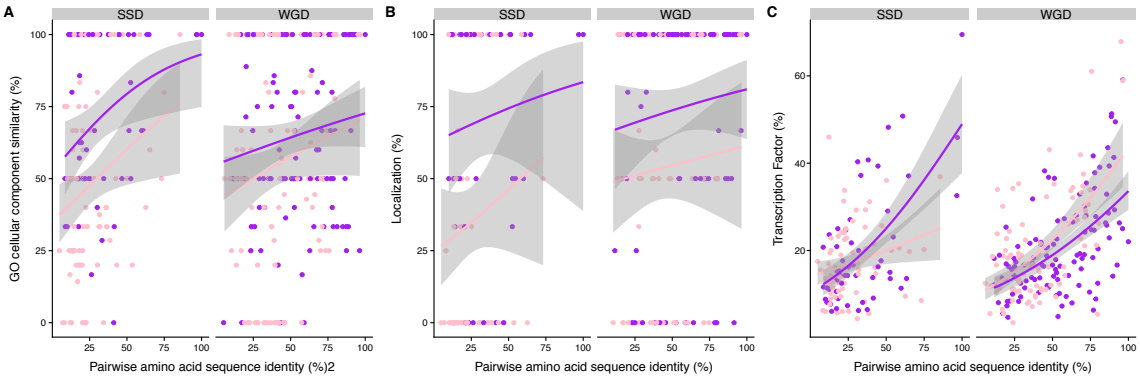

Figure S23: Similarity of regulation between paralogs as a function of their pairwise amino acid sequence identity.
The similarity of co-expression of HM (pink) and HM&HET (purple) pairs was compared while controlling for pairwise amino acid sequence identity for both SSD and WGD. Similarity of co-expression was estimated using **(A)** cellular component similarity GO term, **(B)** similarity of localization and **(C)** similarity of transcription factor binding sites. The regression lines were smoothed using glm method with quasibinomial family.